

SIMULATED INDUCTION  
&  
ITS APPLICATION TO  
BOTANICAL KEY GENERATION

by

E.G. Faulkner, B.E., B.A..

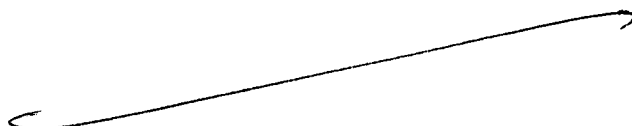
Edwin  
Graeme .

Submitted in fulfilment of the requirements  
for the degree of  
Doctor of Philosophy

University of Tasmania, December 1992

This thesis contains no material which has been accepted for the award of any other higher degree or graduate diploma in any tertiary institution and that, to the best of my knowledge and belief, the thesis contains no material previously published or written by another person, except when due reference is made in the text of this thesis.

*Graeme Faulkner*



## ACKNOWLEDGMENTS

In compiling this thesis I would like to thank my supervisor Phil Collier for the help and encouragement he has given me. I would also like to thank James Alexander for being a sounding board on some psychological issues, Glen McPherson for providing enlightenment on some statistical topics, and Dr. Tony Orchard for allowing the use of specimens and data from the Tasmanian Herbarium. In making these thanks, I would stress that any conclusions made, or any methodology used, are the responsibility of the author alone. Particular thanks are due to John and Amanda Faulkner who allowed me the freedom to pursue this topic, and Dorothy and the late John Faulkner who provided the inspiration and the courage to continue.

## **Abstract**

The dissatisfaction expressed by taxonomists with the results obtained from automatic key-generation methodologies employing deductive logic led to an examination of interactive key generation methodologies employing inductive logic. Philosophical and psychological aspects of induction were examined to ensure that the resulting methodology would be philosophically and psychologically acceptable, and a case was made that such methods would in fact be more widely understood in the community than deductively-based methodologies. The effects of the discussion on the debate about the existence of artificial intelligence and the problems of obtaining rules for expert systems were noted, and a computerised methodology implemented. The results of applying this methodology to Tasmanian data obtained from measurements of specimens of the *Acaena* complex and *Danthonia* genus were compared with several competing methodologies, namely clustering, neural networks, discriminant analysis, a paper-based key produced by a domain expert and entropy-based methodologies. The results obtained were either similar or superior to the competing methodologies; perhaps because the methodology implemented combined the strengths of each of the participants, i.e. the tireless calculating ability of the computer with the background knowledge and common sense of the domain expert. With some types of data, the methodology was also less computationally intensive than some competing methodologies.

## **CONTENTS - SUMMARY**

Abstract	3
Contents - Summary	4
Contents - Sections	5
Contents - Figures	14
Contents - Tables	16
Contents - Equations	21
Introduction	22
Artificial Intelligence and the Development of Human Induction	27
Background to Computer Simulation of Induction	83
A Statistical Approach to Inductive Categorisation	118
Inductive Categorisation Implementation	174
Inductive Categorisation, Dendrograms, and Botanical Data	187
Inductive Categorisation; Key Construction	202
Future Work	249
Conclusions	250
Reference List	252
Appendix A: Clustering Methodology and Categorisation.	285
Appendix B: Neural Network Methodology and Categorisation.	314
Appendix C: Voting Methodology and Categorisation.	351
Appendix D: Discriminant Analysis and Categorisation.	359
Appendix E: Validity of Data Used.	367



## **Table of Contents — Sections**

Introduction	22
Artificial Intelligence and the Development of Human Induction	27
1.1 The Use of Induction in Reasoning	27
1.1.1 Purpose of Induction	28
1.1.1.1 Inductive Classification is not Absolute	28
1.1.1.2 Complete enumeration versus partial enumeration	29
1.1.2 Induction is useful in practice	31
1.1.3 Mill's methods for the use of induction	34
1.1.4 Limitations of Induction	36
1.2 Theories concerning Induction	36
1.2.1 Theories of Helmholtz	37
1.2.2 Theories of Dewey	38
1.2.3 Theories of Rowe	38
1.2.4 "Gestalt" versus "Information Processing" Theories	38
1.2.5 Theories of Piaget	39
1.2.5.1 Period of Sensory-Motor Intelligence	41
1.2.5.2 Stage of Preoperational Thought	42
1.2.5.3 Stage of Concrete Thought	44
1.2.5.4 Propositional or Formal Operations	45
1.3 How widely applicable are these theories?	45
1.3.1 Does everyone achieve deductive logic?	46
1.3.2 Automaticity	49
1.3.3 Automaticity and Induction	53
1.3.4 Limitations of Expert systems	54
1.4 Simulation of Induction	54
1.4.1 What can be simulated?	55
1.4.2 Relationship between Induction and Intelligence	56
1.4.3 Advantages of Induction	61
1.5 Controversy about existence of Artificial Intelligence	63
1.5.2 Beliefs taken as axioms, and their consequences.	64
1.5.2.1 Determinists and Artificial Intelligence	67
1.5.2.2 Libertarians and Artificial Intelligence	68

## Table of Contents — Sections

1.5.2.2.1	Christian Libertarians	72
1.5.2.2.2	Aristotelian Libertarians	74
1.5.2.2.3	Pantheist Libertarians	75
1.5.2.2.4	Parallelist and Epiphenomenalist Libertarians	77
1.5.2.2.5	Summary; Libertarians and Artificial Intelligence	77
1.5.3	Cognitive Modelling and Artificial Intelligence	79
1.5.4	Computer Science and Artificial Intelligence	80
1.6	Summary: induction, humans and expert systems	82
Background To Computer Simulation of Induction		83
2.1	Deriving Rules to Systematise Data	83
2.1.1	Discrete and Continuous Data	84
2.1.2	Data Compression applied to real-numbered characteristics	85
2.1.3	Key Building - A Background History	86
2.1.4	Categorisation of Numeric Data	97
2.1.5	Classification of discrete valued characteristics	97
2.2	Obtaining Rules for use in Expert systems	98
2.2.1	Collecting the Expertise	91
2.2.2	Induction and the Feigenbaum bottle-neck	100
2.2.3	Common Problems with Data of Botanic Origin	106
2.3	'Selecta-key' Specification	117
A Statistical Approach To Inductive Categorisation		118
3.1	Key decisions using a single characteristic	118
3.1.1	Statistics and Inductive Categorisation	119
3.1.2	Tests assuming Parametric Distributions	120
3.1.2.1	Large Sample Tests	121
3.1.2.1.1	Introduction — Properties of a Normal Curve	121
3.1.2.1.2	Distinguishing between two Species	125
3.1.2.1.3	Are the means of two large-sample distributions different?	126
3.1.2.1.4	Separation points in large sample parametric distributions	128

## Table of Contents — Sections

3.1.2.1.5 Distinguishing between many large-sample distributions	134
3.1.2.1.6 Splitting Points and Multiple Distributions	136
3.1.2.1.6.1 Method 1 — 'Grouping'	137
3.1.2.1.6.2 Method 2 — 'Individual Difference'	138
3.1.2.1.7 Type 1 errors and Decision Keys	142
3.1.2.3 Small Sample Parametric Tests	145
3.1.2.3.1 Introduction	146
3.1.2.3.2 Difference between means, small sample parametric distributions	146
3.1.2.3.3 Choosing a splitting point, with small sample distributions	148
3.1.2.3.4 Distinguishing between many small-sample distributions	148
3.1.3 Non-Parametric Tests	148
3.1.3.1 Introduction to non-parametric tests	149
3.1.3.1.1 Sign Test	149
3.1.3.1.2 U Test	149
3.1.3.1.3 Randomisation Tests — Introduction	150
3.1.3.1.4 Randomisation Tests — Advantages & Disadvantages	150
3.1.3.2 Randomisation Tests — Possible method of use	152
3.1.3.2.1 Distinguishing between two non-parametric distributions	153
3.1.3.2.2 Randomisation Tests — Minimum group sizes	157
3.1.3.2.3 Approximate Randomisation Tests	157
3.1.3.2.3.1 Approximate Randomisation tests for large groups	158
3.1.3.2.3.2 Approximate Randomisation tests for small groups	160
3.1.3.2.4 Randomisation Tests — Comparison with parametric tests	160
3.1.3.2.5 Randomisation Tests — Are keys needed?	161

## Table of Contents — Sections

3.1.3.2.6 Randomisation Tests — Approach adopted	162
3.1.3.2.7 Randomisation Tests — Splitting Point Selection	162
3.1.3.2.8 Randomisation Tests — Distinguishing between many distributions	162
3.1.3.3 Randomisation Tests — Summary	163
3.2 Error Correction — Use of Multiple Characteristics	163
3.2.1 Choice of the best single characteristic to use	163
3.2.2 Construction of the key	165
3.2.2.1 Key construction — Expert aided by Selecta-key	165
3.2.2.2 Key construction — Automatic	165
3.2.3 Use of more than one characteristic per decision	167
3.2.3.1 Error Correction using Multiple Characteristics	167
3.2.3.2 Multiple Characteristics and Selecta-key	169
3.3 'Voting' Methodology	172
3.4 Summary — Statistical Methods	173
Inductive Categorisation Implementation	174
4.1 First prototype of Selecta-key	174
4.2 Second prototype of Selecta-key	175
4.3 Third prototype of Selecta-key	177
4.4 Simplified Inductive Classification (Voting).	182
4.5 Checking for Outliers	182
4.6 Aristotelian Neural Net Simulator	183
4.7 Ancillary programs	184
Inductive Categorisation, Dendrograms, and Botanical Data	187
5.1 Botanic Dendrograms — Limitations of the Selecta-key approach.	187
5.1.1 Dendrogram types — Genealogical Dendrograms	188
5.1.2 Dendrogram types — Cladistic Dendrograms	190

## Table of Contents — Sections

5.1.3 Dendrogram types — Identification	
Dendrograms	190
5.2 Requirements of Botanic Data	191
5.2.1 Data Requirements for Comparison of	
Methodologies for the Identification of Botanic	
Specimens	192
5.2.2 Data Requirements for the accurate	
identification of Botanic Specimens	194
5.3 An examination of Data Sets for use in Methodology	
Comparison	195
5.3.1 Choice of data for suitability for use in	
methodology comparisons.	195
5.3.2 An examination of data sets for use in accurate	
botanic specimen identification	196
5.4 Training and Test Sets of Data	198
5.5 Summary	200
Inductive Categorisation & Key Construction	202
6.1 Comparison of Results obtained from Selecta-key	
and 1 <sup>st</sup> Class	203
6.1.1 Key construction — Selecta-key and 1 <sup>st</sup> Class	203
6.1.2 Alternate Key construction — Selecta-key and	
1 <sup>st</sup> Class	206
6.2 Comparing Selecta-key's <i>Acaena</i> and <i>Danthonia</i>	
Keys with existing keys.	210
6.2.1 Results obtained using Selecta-key with the	
<i>Acaena</i> data	211
6.2.1.1 <i>Acaena</i> Data, Collier's Summary Key	211
6.2.1.2 <i>Acaena</i> Data, Quinlan's C4.5 Algorithm	211
6.2.1.3 <i>Acaena</i> Data, Orchard's Key	212
6.2.1.4 <i>Acaena</i> Data, Selecta-key Imitation of	
Orchard's Key	214
6.2.1.5 <i>Acaena</i> Data — Key derived from Selecta-	
key	215
6.2.2 Results obtained using Selecta-key with the	
<i>Danthonia</i> data	217
6.2.2.1 <i>Danthonia</i> data — Collier's Key	217

## Table of Contents — Sections

6.2.2.2 <i>Danthonia</i> data — Key derived from Selecta-key	221
6.3 Computer time used	225
6.4 Statistical-only versus statistical plus randomisation runs	226
6.5 Alternative Methodologies; Implementation and Test Runs	227
6.5.1 Alternative Methodologies — Cluster Analysis	228
6.5.1.1 Purpose of Cluster Analysis	228
6.5.1.2 Natural Number of Clusters	229
6.5.1.3 Rate of identification using Cluster Analysis	229
6.5.1.4 Summary of results using Cluster Analysis	230
6.5.2 Alternative Methodologies — Neural Nets	233
6.5.2.1 Species Identification, Aristotelian neural net	233
6.5.2.2 Species Identification, non-Aristotelian neural net	234
6.5.2.3 Neural net summary	236
6.5.3 Alternative Methodologies — Voting	237
6.5.3.1 Discussion of Voting Methodology	237
6.5.3.2 Results obtained by use of the Voting Methodology	238
6.5.3.3 Summary — Voting Methodology	238
6.5.4 Alternative Methodologies — Discriminant Analysis	239
6.5.4.1 Parametric Discriminant Analysis	239
6.5.4.1.1 Parametric methodology employed	239
6.5.4.1.2 Results obtained from the parametric discriminant analysis	240
6.5.4.2 Non-Parametric Discriminant Analysis	241
6.5.4.2.1 Non-parametric methodology employed	241
6.5.4.2.2 Results obtained from the non- parametric discriminant analysis	241
6.5.5 Summary — Alternative Methods	242
6.6 Discussion of Results	243

## Table of Contents — Sections & Appendices

6.7 Summary of Results	248
Future Work	249
Conclusions	250
Reference List	252
<hr/>	
Appendix A Clustering Methodology	285
A.1 Discussion on Clustering.	285
A.1.1 Finding the number of clusters	286
A.1.2 Poorly and well separated clusters	287
A.1.3 Clustering limitations involving incomplete data	290
A.2 Results obtained using Clustering Methodology	290
A.2.1 Preliminary results using Ward's method plus discriminant analysis with the <i>Acaena</i> and <i>Danthonia</i> data.	291
A.2.2 The number of clusters in the <i>Acaena</i> and <i>Danthonia</i> data.	296
A.2.3 Full clustering results using the <i>Acaena</i> data	298
A.2.4 Full clustering results for the <i>Danthonia</i> data	307
A.3 Summary	313
Appendix B: Neural Networks	314
B.1 Can the human brain be imitated?	314
B.1.1 Attempts to imitate the brain	315
B.1.2 Are rules necessary?	316
B.1.3 Should imitations start bottom-up or top-down?	317
B.1.4 Imitating the brain's parallel processing	317
B.1.5 How does the brain work?	319
B.1.5.1 The black box approach	319
B.1.5.2 The divide-and-conquer approach	320
B.1.5.3 Applying a combined approach	320
B.2 Modelling the Neuron	321
B.3 Joining the model neurons into a network — brief summary	324
B.3.1 Hopfield nets	326
B.3.2 Hamming nets	326
B.3.3 Bi-directional Associative Memory nets	326

## Table of Contents — Appendices

B.3.4 Carpenter/Grossberg classifiers	326
B.3.5 Kohonen nets	327
B.3.6 Neocognition nets	327
B.3.7 Multi-layer perceptron nets	327
B.3.7.1 A 3-layer perceptron net	328
B.3.7.2 Ability of a 3-layer perceptron net to generalise	330
B.3.7.3 More than 3 layers in a perceptron net?	332
B.4 Some Neural Net Theory	333
B.5 Implementation Issues	338
B.6 Results Obtained	339
B.6.1 Training and Test Data Sets	339
B.6.2 Results obtained from Neural Net runs	339
B.6.2.1 Data Treatment	340
B.6.2.2 Experiences with the Aristotelian Net	340
B.6.2.3 Results Obtained with non-Aristotelian Net	342
B.7 Discussion	349
B.8 Summary	349
<b>Appendix C: Voting Methods</b>	<b>351</b>
C.1 Voting Methodology	351
C.1.1 Detail of Methodology	351
C.1.2 Implementation	352
C.2 Results from Voting Methodology	353
C.2.1 Treatment of Data	353
C.2.2 Results from Data	353
C.3 Discussion of Methodology	355
C.3.1 Discussion of Results	356
C.3.2 Advantages and Disadvantages	357
C.4 Summary	358
<b>Appendix D: Discriminant Analyses</b>	<b>359</b>
D.1 Discussion of the Methodologies employed for the Discriminant Analyses	359
D.1.1 Detail of Methodologies	359
D.1.2 Implementation	360
D.2 Results obtained by applying Discriminant Analysis Methodologies	360
D.2.1 Treatment of Data	361



## **Table of Contents — Appendices**

D.2.2 Results from Data	361
D.2.2.1 Results from a parametric methodology	361
D.2.2.2 Results from a non-parametric methodology	363
D.2.3 Discussion of Results	365
D.3 Advantages and Disadvantages	366
D.4 Summary	366
Appendix E: Data Used	367
E.1 Origins of the Data	367
E.2 Types of Data used	367
E.3 Suitability of the Data Characteristics used	369
E.4 Consistency of Data	376
E.4.1 Form of the data	376
E.4.2 Presence of Outliers	377
E.4.2.1 Examination for possible Data Entry errors	378
E.4.2.2 Examination for possibly anomalous specimens	380
E.5 Summary	382

## Table of Contents — Figures

Figure 1 — Examples of inductive reasoning problems	58
Figure 2 — Examples of true and false figural analogies	59
Figure 3 — An IQ test analogical reasoning problem	60
Figure 4 — Normal or Gaussian Probability Curve	122
Figure 5 — Area under Normal Probability Curve	124
Figure 6 — Distributions exhibiting separation	126
Figure 7 — Distributions exhibiting separation	128
Figure 8 — Overlapping distributions	130
Figure 9 — Overlapping distributions	131
Figure 10 — Scores on an Arithmetic Reasoning Test	132
Figure 11 — Incompletely separated distributions	134
Figure 12 — Multiple distributions	135
Figure 13 — Multi-modal distributions with splitting point S1	137
Figure 14 — Multi-modal distributions with small distributions grouped about splitting point S1	138
Figure 15 — Matrix of confidence levels	139
Figure 16 — Splitting point S2 chosen	140
Figure 17 — S3 splitting point chosen	140
Figure 18 — Splitting point S4 chosen	141
Figure 19 — $x_{split}$ in a distribution	176
Figure 20 — A cladogram showing Hennig's definition of a relationship	189
Figure 21 — <i>Acaena</i> key produced by the Selecta-key system	204
Figure 22 — The summary decision key from Collier's Figure 6	204
Figure 23 — <i>Acaena</i> classification key, constructed without flower data	207
Figure 24 — Key from 1 <sup>st</sup> Class, using no-flower, no-fruit <i>Acaena ovina</i> data	208-209
Figure 25 — The decision key produced by the C4.5 algorithm	212
Figure 26 — Orchard's Key to the Australian Species and varieties of <i>Acaena</i> .	213
Figure 27 — Selecta-key key for <i>Acaena ovina</i> taxa	216
Figure 28 — Collier's <i>Danthonia</i> key	218-220

## Table of Contents — Figures

Figure 29 — Selecta-key <i>Danthonia</i> key	222-225
Figure 30 — Species-specific plot of Fisher's Iris data, clustered by Ward's method	288
Figure 31 — Cluster-specific plot of Fisher's Iris data, clustered by Ward's method	289
Figure 32 — Cluster-specific plot of the <i>Acaena</i> data, clustered by Ward's method	293
Figure 33 — Taxa-specific plot of the <i>Acaena</i> data, clustered by Ward's method	294
Figure 34 — Cluster-specific plot of the <i>Danthonia</i> data, clustered by Ward's method	295
Figure 35 — Diagrammatic representation of a Neuron	321
Figure 36 — McCulloch-Pitts model of a Neuron	322
Figure 37 — Main Types of Neural Networks	325
Figure 38 — Three-layer perceptron net	328
Figure 39 — Binary image of a character	329
Figure 40 — Diagrammatic representation of an input- layer Neuron	334
Figure 41 — Diagrammatic representation of a middle- layer Neuron	334
Figure 42 — Shape of a Sigmoid function	335
Figure 43 — Diagrammatic representation of an output- layer Neuron	335
Figure 44 — Delta rule weight changes	336
Figure 45 — Shape of derivative of a sigmoid function	337

## Table of Contents — Tables

Table 1 — A possible leaf length distribution	103
Table 2 — Area between confidence limits	124
Table 3 — Mean and standard deviation	132
Table 4 — Percentage correctly classified after successive questions	144
Table 5 — Observed value of leaf length	154
Table 6 — Randomisation of two <i>Acaena Ovina</i> measurement groups	155
Table 7 — Minimum Group Size for Randomisation Test	157
Table 8 — Binary representation of multiple characteristic decision	170
Table 9 — Binary representation of sample specimens	170
Table 10 — t test results	205
Table 11 — t test results	206
Table 12 — <i>Acaena</i> classification rate obtained by use of Collier's key	211
Table 13 — <i>Acaena</i> classification rate obtained by use of the C4.5 key	212
Table 14 — <i>Acaena</i> classification rate obtained by use of Orchard's key	214
Table 15 — <i>Acaena</i> classification rate obtained by use of "imitation" key	215
Table 16 — <i>Acaena</i> classification rate obtained by use of Selecta-key key	216
Table 17 — <i>Danthonia</i> classification rate obtained by use of Collier's key	220
Table 18 — Classification rate, Selecta-key <i>Danthonia</i> key	225
Table 19 — Relative running times of Selecta-key and ID3 algorithms	226
Table 20 — Rate of identification of <i>Acaena</i> data using Cluster Analysis	231
Table 21 — Rate of identification of <i>Danthonia</i> data using Cluster Analysis	232
Table 22 — Classification rate, Aristotelian neural net method using <i>Danthonia</i> data	233

## Table of Contents — Tables

Table 23 — Average classification rate, non-Aristotelian neural net method using <i>Acaena</i> data	235
Table 24 — Average classification rate, neural net method using <i>Danthonia</i> data	236
Table 25 — Average classification rate, Voting methodology (first choice only)	238
Table 26 — Average classification rate, Voting methodology (first two choices only)	238
Table 27 — Average classification rate, discriminant analysis using the <i>Danthonia</i> data	240
Table 28 — Average classification rate using the <i>Acaena</i> data (excluding the specimens having incomplete data)	240
Table 29 — Average classification rate using the <i>Acaena</i> data (including the specimens having incomplete data)	241
Table 30 — Average classification rate using the <i>Danthonia</i> data, Epanechnikov's kernel methodology	242
Table 31 — Ranges of classification rates using the <i>Danthonia</i> data	242
Table 32 — Ranges of classification rates using the <i>Acaena</i> data	243
Table 33 — Classification of <i>Danthonia</i> data	244
Table 34 — Classification of <i>Acaena</i> data	245
Table 35 — Key to the <i>Acaena</i> data	299
Table 36 — Density Linkage Cluster Analysis for the <i>Acaena</i> data	301
Table 37 — Two-stage Density Linkage Cluster Analysis for the <i>Acaena</i> data	302
Table 38 — Single Linkage & Ward's Cluster Analyses for the <i>Acaena</i> data	304
Table 39 — Average, Complete, Centroid and EML Cluster Analyses for the <i>Acaena</i> data	305
Table 40 — Flexible, McQuitty and Median Cluster Analyses for the <i>Acaena</i> data	306
Table 41 — Key to the <i>Danthonia</i> data	308

## Table of Contents — Tables

Table 42 — EML and Ward's Cluster Analyses for the <i>Danthonia</i> data	309
Table 43 — Average, Complete, McQuitty and Flexible Cluster Analyses for the <i>Danthonia</i> data	310
Table 44 — Density and Two-stage Density Cluster Analyses for the <i>Danthonia</i> data	311
Table 45 — Centroid, Single Linkage and Median Cluster Analyses for the <i>Danthonia</i> data	312
Table 46 — Classification rate, Aristotelian Neural net method, using <i>Danthonia</i> data.	341
Table 47 — Classification rate, non-Aristotelian Neural net method using <i>Acaena</i> data; network configuration = 31 input nodes, 63 hidden nodes and 11 output nodes	343
Table 48 — Classification rate, Neural net method using <i>Acaena</i> data; network configuration = 31 input nodes, 41 hidden nodes and 11 output nodes	343
Table 49 — Classification rate, Neural net method using <i>Acaena</i> data; network configuration = 31 input nodes, 21 hidden nodes and 11 output nodes	344
Table 50 — Classification rate, Neural net method using <i>Acaena</i> data; network configuration = 31 input nodes, 11 hidden nodes and 11 output nodes	344
Table 51 — Average Classification rate, non-Aristotelian Neural Net method using the <i>Acaena</i> data	345
Table 52 — Classification rate, non-Aristotelian Neural net method using <i>Danthonia</i> data; network configuration = 41 input nodes, 83 hidden nodes and 19 output nodes	345
Table 53 — Classification rate, Neural net method using <i>Danthonia</i> data; network configuration = 41 input nodes, 63 hidden nodes and 19 output nodes	346
Table 54 — Classification rate, Neural net method using <i>Danthonia</i> data; network configuration = 41 input nodes, 43 hidden nodes and 19 output nodes	346

## Table of Contents — Tables

Table 55 — Classification rate, Neural net method using <i>Danthonia</i> data; network configuration = 41 input nodes, 23 hidden nodes and 19 output nodes	347
Table 56 — Classification rate, Neural net method using <i>Danthonia</i> data; network configuration = 41 input nodes, 13 hidden nodes and 19 output nodes	347
Table 57 — Classification rate, Neural net method using <i>Danthonia</i> data; network configuration = 41 input nodes, 8 hidden nodes and 19 output nodes	348
Table 58 — Average Classification rate, non-Aristotelian method using <i>Danthonia</i> data.	348
Table 59 — Classification rate — First choice; Voting methodology using <i>Danthonia</i> data	354
Table 60 — Total Classification rate — first choice; Voting methodology using <i>Danthonia</i> data	354
Table 61 — First two choices — Voting methodology using <i>Danthonia</i> data	355
Table 62 — Total Classification rate — first two choices; Voting methodology using <i>Danthonia</i> data	355
Table 63 — Classification rate — multivariate normal methodology using <i>Danthonia</i> data	362
Table 64 — Average classification rate — multivariate normal methodology using <i>Danthonia</i> data	362
Table 65 — Total classification rate for completely described specimens — multivariate normal methodology using <i>Acaena</i> data	363
Table 66 — Total classification rate for all specimens — multivariate normal methodology using <i>Acaena</i> data	363
Table 67 — Classification rate — Epanechnikov's kernel methodology using <i>Danthonia</i> data	364
Table 68 — Average classification rate — Epanechnikov's kernel methodology using <i>Danthonia</i> data	365

## Table of Contents — Tables

Table 69 — Correlations between <i>Acaena</i> characteristics — Part 1	371
Table 69 (continued) — Correlations between <i>Acaena</i> characteristics — Part 2	372
Table 70 — Correlations between <i>Danthonia</i> characteristics — Part 1	373
Table 70 (continued) — Correlations between <i>Danthonia</i> characteristics — Part 2	374
Table 70 (continued) — Correlations between <i>Danthonia</i> characteristics — Part 3	375
Table 70 (continued) — Correlations between <i>Danthonia</i> characteristics — Part 4	375



## Table of Contents — Equations

Equation 1 . . . . .	123
Equation 2 . . . . .	123
Equation 3 . . . . .	123
Equation 4 . . . . .	124
Equation 5 . . . . .	127
Equation 6 . . . . .	127
Equation 7 . . . . .	127
Equation 8 . . . . .	129
Equation 9 . . . . .	129
Equation 10 . . . . .	129
Equation 11 . . . . .	130
Equation 12 . . . . .	130
Equation 13 . . . . .	131
Equation 14 . . . . .	147
Equation 15 . . . . .	147
Equation 16 . . . . .	147
Equation 17 . . . . .	147
Equation 18 . . . . .	155
Equation 19 . . . . .	159
Equation 20 . . . . .	159
Equation 21 . . . . .	159
Equation 22 . . . . .	159
Equation 23 . . . . .	160
Equation 24 . . . . .	160
Equation 25 . . . . .	168
Equation 26 . . . . .	169

# SIMULATED INDUCTION & ITS APPLICATION TO BOTANICAL KEY GENERATION

## Introduction

The subject of this thesis is induction, specifically the application of induction and the acceptability of inductively based artificial intelligence methodologies when applied to key generation in the botanic area.

This area contains many challenges to the key designer. It represents a field of expertise in which the statistical methodologies developed in recent years do not fit easily. Traditional key generation in this area has depended on the use of mainly qualitative rather than quantitative characteristics, typically employing phenotypical rather than genotypical attributes.

This thesis briefly examines the philosophical background and historically noted problems of employing inductive methodologies (rather than the deductive methodologies more usually employed in the physical sciences) and notes the consequential philosophical and practical limitations imposed on the subsequent use of the resultant key or expert system produced by the application of inductive methodology.

Before the key or expert system builder can feel comfortable with a tool, he or she must have sufficient background to appreciate the methodology. This thesis briefly argues (from a psychologically inclined developmental view) that the development of inductive reasoning is (given the absence of major developmental handicaps) a part of every person's developmental process, and thus available to every adult human. It is noted, by contrast, that many psychologists believe that deductive reasoning is developed by only a proportion of the

mature human population, and thus has less universality of application than is desirable in a general-purpose tool.

Before the key builder can gain the full benefits of the application of the methodologies of artificial intelligence, it is fundamental that he or she should feel comfortable with the concept of intelligence (or at least a portion thereof) residing within an artificial construct. This thesis argues that the level of comfort with, and acceptance of these concepts will depend largely on the basic axioms and constructs of belief employed in the expert's belief system. Some aspects of belief systems are briefly examined with respect to the algorithmic objection to the existence to artificial intelligence, and the likely effect of those belief systems on the acceptance of the concept of artificial intelligence is postulated.

With this background, a methodology of key construction is suggested which allows the combination of elements of the qualitative expertise of the domain expert with the calculative ability of the computer. This is done by presenting the expert with a list of statistically valid alternative splitting points; separate computational procedures being available for normal & non-normal populations. The inclusion of the expert in the system is felt to be vital, given the qualitative nature and sampling problems typical in much data of botanic origin. Inherent in this methodology is the elimination of the necessity to prune the resulting key, and the ability to produce polythetic as well as the more usual monothetic keys.

The proposed methodology was tested against a variety of traditional techniques including clustering, neural networks, voting, discriminant analysis, entropy-reduction and paper-based keys produced by domain experts. The proposed methodology offered a considerable saving in computational load when producing the initial key, and produced results which were better or comparable with alternate methodologies, particularly when used with data sets which were typical of many botanic data sets in that they were large, poorly separated data sets in which a proportion of the specimens were incompletely described.

The work is described under eight main headings, with alternative methodologies and the validity of the data being examined in a further five Appendices.

- This **Introduction** explains the purpose and plan of this thesis.
- The **Artificial Intelligence and the Development of Human Induction** section defines induction in the relatively well established psychologically based terms of human development. Comparisons are made between human and expert system abilities at various stages of human development. This comparison occurs against the philosophical background of induction, whilst acknowledging that there is a controversy about the existence of artificial intelligence. It then briefly suggests that the type of classification accorded intelligent machines may be predicated by the (often unexamined) philosophical axioms of belief employed by the person making the judgement. It is argued that some sets of axiomatic beliefs make the idea of machine intelligence quite unacceptable, while others put no obstacle in the path of the acceptance of the concept of intelligent machines.
- **Background to Computer Simulation of Induction** examines some of the previous successes and problems in simulating inductive processes by computer.
- **A Statistical Approach to Inductive Categorisation** suggests some methods which could simulate induction in such a way that the results would both be more acceptable and understandable to many human experts, and may be more reliable in the face of missing or mis-classified data.
- **Inductive Categorisation Implementation** discusses implementations of the inductive categorisation algorithms introduced in the previous chapter. The implementation of these inductive categorisation algorithms will be referred to in the rest of this thesis as the Selecta-key programs. This chapter also contains some brief comments on some other necessary programs developed during the course of this study to supplement the Selecta-key programs

- **Inductive Categorisation, Dendrograms, and Botanical Data** comments on the construction of dendritic trees (keys) in botany. Three types of dendrograms are commonly referred to in botanic literature. These types are discussed in the light of the applicability of the Selecta-key approach to their construction. The problems which typically occur in collections of botanic data are also examined, with a view to ensuring that the data employed in the comparisons between Selecta-key and other methodologies adequately represents the problems typical of botanic data collections.
- The section entitled **Inductive Categorisation; Key Construction** examines both the implementation of this statistical approach to inductive categorisation in the Selecta-key programs, and the results obtained by applying entropy-based and Selecta-key systems to key construction for the *Acaena ovina* and *Danthonia*<sup>1</sup> botanical data. It also compares the results of the suggested methodology with previously applied methods, including discriminant analysis, various clustering procedures, two implementations of neural net methodology, and a simpler variation of the Selecta-key methodology referred to as the voting methodology. It concludes with discussion of the comparative results.
- The **Future Work** section summarises briefly the future directions work on this project could take to extend the work and make the results easier for a domain expert to use.
- The **Conclusion** section briefly summarises the ideas proposed and the work achieved in this project.
- The **Reference List** section lists references referred to in the body of this thesis.

---

<sup>1</sup>The *Acaena* data was made up from measurements taken from specimens collected and made available by the Curator of the Tasmanian Herbarium, Dr. A. R. Orchard. He also made *Danthonia* specimens available for measurement. The author wishes to express sincere thanks to Dr. Orchard for his generosity in making this data available, and to stress that any conclusions made about or from the data are the responsibility of the author alone.

- **Appendix A: Clustering Methodology and Categorisation** details results obtained by applying clustering methodology to the *Acaena* and *Danthonia* data, and makes these results available for comparison with the other methodologies considered in this thesis.
- **Appendix B: Neural Network Methodology and Categorisation** examines the background of attempts to imitate a postulated method of information processing in the human brain, looks at some competing architectures, and applies two variations of these methodologies in relation to the inductive categorisation of the *Acaena* and *Danthonia* data.
- **Appendix C: Voting Methodology and Categorisation** introduces a variation of the methodology proposed in this thesis, examines its advantages and disadvantages, and notes the results obtained when this methodology is applied to the *Acaena* and *Danthonia* data.
- **Appendix D: Discriminant Analysis and Categorisation** uses the statistical approach of discriminant analysis to provide results for comparison with the other methodologies.
- In **Appendix E: the Validity of Data Used** is examined, to check if the *Acaena* and *Danthonia* data are anomalous data. Comments on the suitability of the data characteristics chosen to specify the data are made, and the likely consistency of the data is examined, together with checks for outliers and possible data entry errors.

# ARTIFICIAL INTELLIGENCE AND THE DEVELOPMENT OF HUMAN INDUCTION.

Induction is a significant topic in the area of expert systems, which is currently an important commercial application of artificial intelligence. In this section the philosophical and psychological background to the development and application of human induction is examined. Section 1.1 examines the use of one aspect of artificial intelligence, induction, in human reasoning. Section 1.2 notes theories of induction which have been advanced by various authorities. Section 1.3 considers the extent to which these theories are germane to the general human population. Section 1.4 looks at the simulation of induction. Section 1.5 refers to the controversy about the very existence of Artificial Intelligence, and section 1.6 makes some summary conclusions about the differences in the use of induction by human and expert systems.

## 1.1 The Use of Induction in Reasoning

Before examining the development of human induction, it is useful to review some philosophical opinions regarding the use of induction, as it has been claimed that inductive methodology can not be used as a universal panacea; there are some circumstances where the use of induction is acceptable, other circumstances where it should only be used with caution, and yet other circumstances when it is inadvisable to use inductive methodologies at all. In the following pages section 1.1.1 examines the purpose of induction, noting the difference between complete and partial enumeration of the possible options. Section 1.1.2 notes that the justification of the use of induction is that, despite the fact that inductive classification is not usually absolute, it remains a useful and vital satisficing methodology. Section 1.1.3 notes the philosopher Mill's methods for the use of induction, and section 1.1.4 notes the limitations in the applicability of inductive inference to the general population.

### 1.1.1 Purpose of Induction

English and English define an inductive test as 'one in which the task is to derive a principle from a number of particular examples'.<sup>1</sup> Chaplin concurs.<sup>2</sup> Hence by implication, the application of induction assumes there is some underlying pattern or theory implicit in the raw data. The goal of the inductive process is to reveal and document these patterns.

#### 1.1.1.1 Inductive Classification is not absolute

Chalmers illustrates the basic principle of induction by stating:-

If a large number of As have been observed under a wide variety of conditions, and if all those observed As without exception possessed the property B, then all As have property B.<sup>3</sup>

However Kant warns:-

"Experience teaches us ... that a thing is so and so, but not that it cannot be otherwise."<sup>4</sup>

Hence it is worth emphasising that an inductive classification is not absolute, as:-

strictly speaking, one should not say: "This animal is a sparrow", but: "This animal is more (or less) sparrow than this or those animals", just as we would say of an object that it is "more (or less) brown than ...".<sup>5</sup>

Bloomfield makes a similar point;

---

<sup>1</sup>English, Horace B., and English, Ava C., *A Comprehensive Dictionary of Psychological and Psychoanalytic Terms*, Longmans, Green and Co., New York, 1958, p. 260.

<sup>2</sup>Chaplin, J. P., *Dictionary of Psychology*, Dell Publishing Co., New York, 1975, p. 256.

<sup>3</sup>Chalmers, A. F., *What is this thing called Science*, Second Edition, University of Queensland Press, St Lucia, 1982, p. 5.

<sup>4</sup>Quoted by Aune p. 88; see: Aune, Bruce, *Knowledge of the External World*, Routledge, London, 1991.

<sup>5</sup>Piaget, *The Child's Conception of Physical Causality*, p. 298.



for example, on the basis of observing the attribute "whiteness" among several different members of the swan species one might infer that all swans were white.<sup>1</sup>

Mackie uses a similar illustration.<sup>2</sup>

The potential problem that can occur if inductive reasoning is associated with incomplete enumeration was starkly illustrated by Bertrand Russell, who gave the example of a turkey which inductively assumed that he was always fed at 9:00 a.m.; after all, the regular 9:00 a.m. feed occurred under all conditions, regardless of day of week, fine or inclement weather, number of other occupants in his cage, and any other variant he observed. This inductively derived rule worked well for him until Christmas day when at 9:00 a.m. his throat was cut and he was eaten for dinner.<sup>3</sup>

Russell's response to this problem

was to say that "induction as such" cannot be justified because "it can be shown to lead to falsehood as often as truth."<sup>4</sup>

### 1.1.1.2 Complete enumeration versus partial enumeration

Induction was formulated by Aristotle, and by derivation means *a leading on* (by contrast, the derivation of *deduction* means *a leading down from*). Aristotle intended his logic to be *Induction by Complete Enumeration*, but since complete

---

<sup>1</sup>Bloomfield, Brian P., 'Capturing expertise by rule induction,' in *The Knowledge Engineering Review*, Cambridge University Press, Vol. 2, No. 1, March 1987, p. 56.

<sup>2</sup>Mackie, J. L., 'The Paradox of Confirmation', in *Probabilities, Problems and Paradoxes*, Luckenbach, Sidney A., (Ed.), Dickenson Publishing Company Inc., 1972, pps. 241-252.

<sup>3</sup>*ibid.*, p. 14. Similarly Coady, in his recent book on Testimony, quotes Locke "We might recall the case of the King of Siam discussed by Locke and Hume. As Locke tells it, the King when informed by a certain Dutch ambassador 'that the water in his country would sometimes in cold weather be so hard that men walked on it, and that it would bear an elephant if he were there' replied, 'Hitherto I have believed the strange things you have told me, because I look on you as a sober, fair man: but now I am sure you lie'" Coady, C. A. J., *Testimony*, Oxford University Press, Oxford, 1992, p. 180. The reference to Locke given by Coady is: 'John Locke, *An Essay on Human Understanding*, bk. iv, ch. xv, s. 5.'

<sup>4</sup>Attributed to Russell in Aune, p. 167.

enumeration is rarely possible,<sup>1</sup> induction has long been criticised. Francis Bacon (1561-1626) criticised it as follows:-

The induction which proceeds by simple enumeration is childish; its conclusions are precarious, and exposed to a peril from a contradictory instance; and it generally decides on too small a number of facts, and on those only which are at hand.<sup>2</sup>

Margaret A. Boden comments:-

'Induction' carries overtones of the loose, the shoddy, and the impure, if not of the positively indecent. Even those, like Russell, who defend induction clearly regard it as the poor man's deduction.<sup>3</sup>

Crowson comments on the likely reason for these scathing opinions:<sup>4</sup>

For the natural philosophers, the ultimate test of a scientific law is its predictive power; I think this criteria could well be applied to the generalisations of natural history too. In general, the principles of natural history are of the nature of inductive generalisations—they are not to be established, as many laws of physics have been, by a single crucial experiment, but by the accumulation of a large number of supporting instances. A phenomenon which is no doubt connected with the current eclipse of natural history is the anti-inductive bias of nearly all recently influential philosophers. A favourite word among the English-speaking ones has been 'rigour', which can be translated as 'relying exclusively on strict deductive methods'. Rigour of this sort is not characteristic of natural history, which

---

<sup>1</sup>Dietterich states 'In practice, it is rare for a learning algorithm to have even 50% of the possible training examples available for learning. Similar arguments have been put forward concerning the learning power of back propagation'; see: Dietterich, Thomas G., 'Limitations on Inductive Learning', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishers Inc., San Mateo, U.S.A., 1989. Note that back propagation is discussed in Appendix B of this thesis.

<sup>2</sup>Quoted in Luce, A. A., *Teach Yourself Logic*, English Universities Press, London, 1958, p. 176.

<sup>3</sup>Boden, Margaret A., 'Real World Reasoning', in Cohen, L. Jonathon, & Hesse, Mary (Ed.), *Applications of Inductive Logic*, Clarendon Press, Oxford, 1980, p. 359.

<sup>4</sup>Note that when Crowson refers to 'inductive generalisations' he is assuming partial enumeration; (complete enumeration being almost invariably impossible to achieve in the field of natural history).

is characteristically ignored by these philosophers when looking for scientific illustrations of their theories.<sup>1</sup>

Similarly:-

Scientific explanation is usually described as a *deduction* of a statement describing what has to be explained from *premisses* which include a) general "laws of nature", b) description of "initial conditions." ... He also speaks about "inductive explanations" but, in accordance with our deductive (i.e. Popperian) position, we will not accept this concept into our framework.<sup>2</sup>

Considering the 'anti-inductive bias of nearly all recently influential philosophers' noted above, one may reasonably wonder why induction is used at all.

### 1.1.2 Induction is useful in practice.

However induction is used. Crowson gave an example of the use of inductive prediction in the area of natural history when he wrote:

My generalisation ... was based on the examination of only a few hundred species, and thus might appear as a rather bold piece of induction. I had, in fact, predicated a character for something like 10,000 genera on the basis of its presence in about 200 of them—and so far no exceptions have been brought to light.<sup>3</sup>

Crowson continues:

... in other instances, generalisations of this sort ... have often proved to be subject to some exceptions, however, in cases like

---

<sup>1</sup>Crowson, p. 12.

<sup>2</sup>Kroy, Moshe, *Moral Competence, An Application of Modal Logic to Rationalistic Psychology*, Mouton, The Hague, 1975, p. 44. note that in the original text the quotation above included references which I deleted to assist clarity. The references were: 'K. R. Popper, *The Logic of Scientific Discovery* (Harper, 1959), p.60. (b) C. G. Hempel, *Aspects of Scientific Explanation and other Essays in the Philosophy of Science* (The Free Press, 1965), Pt. 1. '... '(c) I. Sheffler, *The Anatomy of Inquiry* (Knopf, 1967), pps. 25-31'.

<sup>3</sup>Crowson, p. 13.

these, a prediction which proves to be right in only 95 per cent of instances is still worth making.<sup>1</sup>

Simon notes a similar principle relevant in economics when he comments:-

In the face of this complexity the real-world business firm turns to procedures that find good enough answers to questions whose best answers are unknowable. Thus, . . . economic man is in fact a satisficer, a person who accepts "good enough" alternatives, not because he prefers less to more but because he has no choice.<sup>2</sup>

The economic concept of man as a satisficer applies in the cases examined by this thesis. Induction may not prove anything in the sense that *modus ponendo ponens* or *modus tollendo tollens* does, but if it is the best available, its use is a satisficing solution to the situation in hand.

Finally, consider the engineering profession where much construction is based on *codes*. Codes are repositories of recommended practice, necessary because the idealised mathematical models provided by science are generally not adequate to deal with the real-life situations faced by engineers. As an example consider the problems faced by an engineer who wishes to construct in wood. Mathematical models assume a uniform, continuous material, whereas wood is cellular, with properties which vary considerably both across the grain, between trees (even of the same species), and whose strength is effected by the presence or absence of knots. Even the treatment used to dry moisture from the green timber can effect strength, with the effects varying from the outside to the inside of the stack being dried, and between different sizes of timber situated similarly within the stack. To resolve these problems, engineers have historically taken an inductive approach. A large number of sample pieces of wood are tested, and an inductively derived prediction as to the safe strength of all timber is made, based on the small number of samples tested. These predicted strengths are recorded in the recommended code of practice, and used for

---

<sup>1</sup>Crowson, p. 14.

<sup>2</sup>Simon, Herbert A., *The Sciences of the Artificial*, Second Edition, The MIT Press, Cambridge, Massachusetts, 1985, p. 36.

design of timber structures. Similar inductively derived codes of practice are the basis of much of engineering, and are the reason many engineers define their discipline as an art, not a science. The success of this inductive approach can be seen in almost everything artificial which both surrounds and is used by *homo sapiens sapiens*.. Perhaps, as Francis Bacon comments 'Our only hope therefore lies in a true induction'.<sup>1</sup>

The justification of induction in the previous paragraphs may be summed up by saying that, in practice, induction is *useful*.

However modern mathematical theorists and philosophers have gone further, challenging the older ideas about the postulated supremacy of the deductive approach. The experiential mechanism which permits and succours the type of success mentioned above is suggested by Aune, who comments (in a book published last year that 'was certainly stimulated by what can be called the new mathematical inductive logic'):-<sup>2</sup>

According to an influential school of statisticians whose characteristic claim was first enunciated by C. S. Peirce, the basic probabilities needed for experimental inference do not have to be well founded or accurate in some sense. Experimental inference based on Bayes' theorem is self correcting: if one begins with prior probabilities that are not extreme (close to 0 or 1) and continues to update one's probability functions by the rule of conditioning, the effect of one's initial uninferred probabilities will become progressively smaller as one proceeds, so that two people starting out with different basic probabilities and updating their probability functions by successive conditioning involving the same evidential input will eventually agree on the probabilities they ascribe to relevant hypotheses. This claim, which can be demonstrated mathematically,<sup>3</sup>

leads to a situation where:

---

<sup>1</sup>Bacon, Francis, *First Book of Aphorisms*, quoted by Forsyth, R. S. in 'The Evolution of Intelligence', *The Third International Expert Systems Conference*, Learned Engineering, Oxford, 1987, p. 61.

<sup>2</sup>Aune, p. xli.

<sup>3</sup>Aune, pps. 172 - 173. Aune comments that 'Peirce's claim was that "properly conducted inductive research corrects its own premisses"; see Charles Peirce, *Collected Papers*, vol. 5, para 576.'; Aune p. 230.

people who receive the same experiences will naturally move towards a consensus on the probability of causes. The probabilities will be objective in the sense that informed people with adequate experience can agree about them<sup>1</sup>

Aune continues:

It is important to realise that, as regards to the facts of the world, the logic of inductive inference is comparable to that of deductive inference. The latter cannot tell us what, absolutely speaking, is true about the world; it can merely tell us what is true *if* something else is true.<sup>2</sup>

It seems that Aune is, in emphasis, essentially agreeing with the comment made a decade earlier by Boden:-

Inductive reasoning 'as she is spoke' is more worthy of epistemological respect than is commonly allowed by logicians. If one is to take into account the real computational constraints upon real computational systems, then the norms of real — or even artificial — thinking have at least as much right to be treated as normative as do the rules of deductive logic. For rationality cannot in practice do without them.<sup>3</sup>

### 1.1.3 Mill's methods for the use of Induction

Given that induction is useful in practice, the next step is to formalise a method of using it. John Stuart Mill (1806-73) suggested five Methods, here expressed as rules.

The Rule of Agreement.

If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon.<sup>4</sup>

---

<sup>1</sup>Op. cit..

<sup>2</sup>Op. cit..

<sup>3</sup>Boden, 'Real World Reasoning', p. 375.

<sup>4</sup>Mill, John Stuart, *A System of Logic Ratiocinative and inductive*, eighth edition, Longmans, Green and Co., London, 1884, p. 255.

### The Rule of Difference.

If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former, the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.<sup>1</sup>

### The Rule of Agreement and Difference. ('The Joint Method')

If two or more instances in which the phenomenon occurs have only one circumstance in common, while two or more instances in which it does not occur have nothing in common save the absence of that circumstance; the circumstance in which alone the two sets of instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.<sup>2</sup>

### The Rule of Residues.

Subduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents.<sup>3</sup>

### The Rule of Concomitant Variations.

Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation.<sup>4</sup>

Variations of these rules are used in many applications, such as Pople's abductive reasoning in a medical diagnosis system, where a collection of symptoms evokes a hypothesis as to the cause of the symptoms.<sup>5</sup>

---

<sup>1</sup>Ibid., p. 256.

<sup>2</sup>Ibid., p. 259.

<sup>3</sup>Ibid., p. 260.

<sup>4</sup>Ibid. p. 263.

<sup>5</sup>Quoted by Winograd, Terry, 'Computer Programs for Inductive Reasoning', in Cohen, Jonathon and Hesse, Mary (Ed.), *Applications of Inductive Logic*, Clarendon Press, Oxford, 1980, pps. 354-355.

It is important to note that in Pople's system the result is regarded as an *hypothesis* not as a *proven fact*. Much of the past criticism of inductive logic may be attributed to confusion between these two. It is interesting and a little ironic in this context to note that Bloomfield uses the instance of the 'whiteness of the swan'. This argument has been used in many of the older philosophical texts, but in more recent ones written since the discovery of Australia, it has often been replaced by 'the blackness of the raven'.<sup>1</sup> The reason for this is evident as I look out my window in Tasmania; the only swans I can see are black.

#### 1.1.4 Limitations of Induction

The example of the swans clearly illustrates the fundamental limitation of an inductive system; i.e. the conclusion obtained by an inductive system only applies to the classes of data contained in the system's data base. If data pertaining to a data type that is new or unknown to the system is typed in, a false classification may occur, the identification achieved possibly being the one known to the system that is closest to the input data.<sup>2</sup>

Within this limitation, it is suggested that induction is useful.

## 1.2 Theories concerning Induction

Given that induction is useful in assisting to solve some problems, it may well be useful to simulate induction. An inductive procedure could be used to front end an expert system by providing some or all of the classificatory expertise needed by that system. It would also, incidentally, allow the composite induction/expert system combination to meet Schank's

---

<sup>1</sup>Hempel, Carl C., 'Studies in the Logic of Confirmation', in Luckenbach, Sidney A., (Ed.), *Probabilities, Problems and Paradoxes*, Dickenson Publishing Company, California, 1972, pps. 223-230; also Chalmers, p. 14; also Holland, John H., Holyoak, Keith J., Nisbett, Richard E. and Thagard, Paul R., *Induction*, The MIT Press, Cambridge, Massachusetts, 1987, p. 6.

<sup>2</sup> This point is important in the context of induction and the incorporation of inductively derived rules in an expert system. Many expert systems contain no facility to give an indication that a false classification may have been made. By comparison, the first prototype neural net model (used for comparison purposes later in this thesis) was written with this in mind. It is referred to as an "Aristotelian Neural Net" as it assumes Aristotle's principle of complete enumeration, and warns if an example is encountered which is not in the net's experiential data base. Hence in "recognition" mode it gave an "unknown" response if it detected an input pattern it had not previously encountered, rather than giving a "nearest match" response which could possibly be erroneous. For further discussion, see Appendix B of this thesis.



definition of an A.I. process as, 'the science of endowing programs with the ability to change themselves for the better as a result of their own experiences'.<sup>1</sup>

To understand what is being simulated, some background discussion of induction and human cognition is necessary.

The process by which humans solve problems has long been an area of interest to philosophers, who speculated about the processes involved. In the following paragraphs, sections 1.2.1, 1.2.2, 1.2.3 and 1.2.5 outline the approach to induction taken by Helmholtz, Dewey, Rowe and Piaget, respectively. Section 1.2.4 notes the fundamental difference between the approaches of the 'indivisible whole' and 'divide-and-conquer' theorists. The approach of Piaget is examined in somewhat greater detail than is the case of the other theorists, and through his approach induction is placed in the context of human cognitive development.

### 1.2.1 Theories of Helmholtz

Helmholtz (1894) suggested that the inductive process, where a problem P exists, is:-<sup>2</sup>

- 1) Investigation of P in all directions;
- 2) Not consciously thinking about P;
- 3) Appearance of 'happy idea'.

This approach would seem to accord with the Gestalt theory that responses are 'properties of the whole ... and are not derived by summation of its parts. ... The notion of "parts" with attributes of their own, independently of the whole, is held to be misleading'.<sup>3</sup>

---

<sup>1</sup>Schank, Roger, *A.I. Magazine*, Winter/spring 1983, quoted by Amoliar, Stephen W., *Induction: Processes of Inference, Learning and Discovery*, IEEE Expert, Computer Society of the IEEE, USA, Fall 1987, p.92.

<sup>2</sup>Rowe, Helga A. H., *Problem Solving and Intelligence*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1985, pps. 120, 121.

<sup>3</sup>English et. al., p.225. For more comments regarding gestalts, see section 1.1.4 of this thesis.

### 1.2.2 Theories of Dewey

Dewey (1910) took a different approach, suggesting that the process of induction involved:- <sup>1</sup>

- 1) Felt difficulty;
- 2) Location and definition;
- 3) Possible solutions;
- 4) Reasoning;
- 5) Acceptance or rejection.

Dewey's model of problem solving allows reasoning, and the possible subdivision of the problem.

### 1.2.3 Theories of Rowe

Rowe reviews other theoretical approaches by Wallas, Rossman, Young, Polya, Hutchison, Mawardi, Osborn, Skemp, Newell and Simon, Johnson, Anderson and Sternberg.<sup>2</sup> She then postulates what she calls a root model, with each problem being broken into smaller parts which may be solved independently.<sup>3</sup>

### 1.2.4 "Gestalt" versus "Information Processing" Theories

The Gestalt or 'indivisible whole'<sup>4</sup> approach exemplified by Helmholtz and others, and the 'information processing, divide and conquer' approach taken by Dewey, Rowe and others, have both competed for researchers' attention as they have examined human development.

Critics of knowledge engineering such as the Dreyfuses argue that experts do not use rules but, rather, intuitive processes

---

<sup>1</sup>Rowe, *ibid.*

<sup>2</sup>Rowe, pps. 119 - 126.

<sup>3</sup>The 'root model' is similar in form to the 'hierarchical tree' model used in inductive inference. It is discussed in Rowe, pps. 127-129.

<sup>4</sup>English and English in p. 225 define Gestalt Theory as 'the systematic position that psychological phenomena are organised, undivided, articulated wholes or **gestalts**. The properties of a gestalt are properties of the whole as such and not derived by summation of its parts. Conversely, the parts derive their properties from their membership in the whole. The notation of "parts" with attributes of their own, independent of the whole, is held to be misleading.'; (the emphasis and punctuation are as used by English and English).

built up through experiences which are stored in the expert's memory<sup>1</sup>

This places the Dreyfuses close to the Gestalt view of humanity, with knowledge part of a largely indivisible whole, and hence common sense being *inductively* derived from the knowledge base of previous experiences.

By contrast, many researchers in the artificial intelligence area accept (at least implicitly) the information processing approach that simulation of portions of intelligent human behaviour is possible, and very often researchers holding this view use primarily *deductive* logic in their investigations.

The implications of these disparate views on the inductive process will be discussed later in this chapter, after the psychologically based views of Piaget have been considered.

### 1.2.5 Theories of Piaget

Piaget studied children's cognitive development from 1921 to 1980, and his subsequent work led to the development of his theory of *genetic epistemology*, which is probably the most unified theory of human intellectual and cognitive development. Writing in 1992, Anderson comments:

It is safe to say that, as yet, nothing has replaced Piagetian theory as a general theory of cognitive development.<sup>2</sup>

For this reason, we will look at the cognitive theory of the development of induction mainly from the point of view of Piaget's ideas.

Whereas Wechsler writes:

Intelligence . . . is the aggregate or global capacity of the individual to act purposefully, to think rationally and deal effectively with his environment.<sup>3</sup>

---

<sup>1</sup>Bloomfield, pps. 59-60.

<sup>2</sup>Anderson, Mike, p. 115.

<sup>3</sup>Wechsler, David, *The Measurement and Appraisal of Adult Intelligence*, The Williams & Wilkins Company, Baltimore, U.S.A., 1958. p. 7.

Watson and Lindgren comment that Piaget makes no sharp distinction between thought and intelligence, for him they are both aspects of the same central cognitive process.<sup>1</sup> Anderson agrees.<sup>2</sup> Bee comments that this central cognitive process has two essential components, *adaption* and *organisation*.<sup>3</sup> *Adaption* is further composed of *assimilation* and *accommodation*; where *assimilation* is the process of taking in and incorporating happenings and experiences into a person's existing repertoire of stratagems and systems, and *accommodation* is the twin process of adapting the concept or idea to conform with what has been taken in.<sup>4</sup> *Organisation* of experience into the person's schema includes integrating experiences from several senses and the application of induction to classify and group impinging stimuli into sets of systems.

All children apply these processes. Their ability to do so varies with age. The age at which they apply them may also be modified by external or internal factors, (e.g. genetic or social disadvantage, which can modify both mental<sup>5</sup> and physical development<sup>6</sup>). Nash comments that the Piagetian model splits human cognitive development into four stages.<sup>7</sup>

#### Lawler notes

In Piaget's work, a stage is a period of time in which a mind deals in a characteristic fashion with problems encountered in all domains;<sup>8</sup>

---

<sup>1</sup>Watson, Robert I., & Lindgren, Henry Clay, *Psychology of the Child*, John Wiley & Sons, Inc., New York, 1959, p. 164.

<sup>2</sup>'individual differences in intelligence are a property of thought', Anderson, Mike, p. 212.

<sup>3</sup>Bee, Helen, *The Developing Child*, Harper & Row, New York, 1978, p. 197.

<sup>4</sup>Piaget, Jean, *The Origin of Intelligence in the Child*, Penguin Books, Harmondsworth, England, 1983, pps. 160-166.

<sup>5</sup>Bee, Helen, *Social Issues in Developmental Psychology*, Harper and Row, New York, 1978, pps. 237, 311 - 316; see also Anderson, Mike, pps. 86 - 87.

<sup>6</sup>Gardner, Lytt I., *Deprivation in Dwarfism*, in *The Nature and Nurture of Behaviour, Readings from the Scientific American*, W. H. Freeman and Company, San Francisco, 1973, pps. 101 - 107; see also Strickberger, Monroe W., *Genetics*, The Macmillan Company, New York, 1968, pps. 468-473.

<sup>7</sup>Nash, John, *Developmental Psychology*, Prentice/Hall International Inc., London, England, 1973, pps. 361 - 363.

<sup>8</sup>Lawler, R. W., *Computer Experience and Cognitive Development*, Ellis Horwood Limited, West Sussex, England, 1985, p. 73.

but Lawler (influenced by Seymour Papert) prefers the description

A stage is no more than the achievement of a common level of performance across those clusters of cognitive structures which are potentially able to be influenced by a specific cognitive ideal.<sup>1</sup>

Although with *décalage*<sup>2</sup>, these stages can occur at various ages, average development places the stages within the approximate age ranges given below.<sup>3</sup>

#### 1.2.5.1 *Period of Sensory-Motor Intelligence.*

This period lasts from birth to approximately two years of age.

Some authorities comment 'there is a high degree of predetermination, or "hard-wiring," in the mammalian brain'.<sup>4</sup> Similarly, although the expert system is empty of knowledge initially, it is often "hard-wired" as to the *type* of knowledge that it can accept.

As the child grows in experience, these inborn responses are gradually freed from the eliciting stimuli, and in adults there appears to be much less that is hard-wired. Here there is a fundamental difference with expert systems. Expert systems rarely can use experience which occurs after their initial "education". Even if they can, the author is not aware of any in which the *type* of knowledge that is useable can be varied by the expert system itself as a result of its interaction and continuing experience with the "outside world".

Towards the end of this period, the child develops the concepts of objects as stable objects. However Bower, Watson and Lindgren and Nash comment that there is no evidence early in this stage that the child recognises the continued existence of an

---

<sup>1</sup>ibid..

<sup>2</sup>Horizontal *décalage* refers to the processes of child development where the order of development in children is claimed to be parallel but the steps in which that development is claimed to take place may be disjunctive with respect to time.

<sup>3</sup>Nash, pps. 361 - 363.

<sup>4</sup>Thompson, Richard F., *The Brain*, W. H. Freeman & Company, New York, 1985, p. 249.

object outside the child's perceptual field. 'Out of sight' is apparently completely 'out of mind'.<sup>1</sup>

It is interesting to note that the average expert system has no knowledge of data apart from that which it 'perceives', and may perhaps be compared to this aspect of this stage.<sup>2</sup>

### 1.2.5.2 *Stage of Preoperational Thought.*

This occurs between the approximate ages of two and seven years of age. The ability to classify objects into concepts starts to appear. This is the start of an inductive ability.

However, the cognitive ability to deal with those classifications has not yet developed. Nash quotes one of Piaget's examples,

a child walking through a wood sees several snails; he does not know whether he sees the same snail repeatedly or a different snail each time; the distinction is, in fact, meaningless to him. The concepts of 'snail in general' and 'this snail in particular' are not yet learned.<sup>3</sup>

Thought at this stage is intuitive and irreversible. A child shown a ball of clay rolled out into a sausage shape will be likely to say that there is 'more clay there because it is longer'.<sup>4</sup> If the clay is then rolled back into the original ball the child is unable

---

<sup>1</sup>Bower, T. G. R., *A Primer of Infant Development*, W. H. Freeman & Company, San Francisco, 1977, p. 110; Watson & Lindgren, p. 165; Nash, p. 361. See also Goldman-Rakic, Patricia S., 'Working Memory and the Mind', *Scientific American*, Vol. 267 No. 3, September 1992, p. 74 where she confirms Bower's observation, and comments that a similar behaviour is displayed by monkeys whose prefrontal regions have been surgically ablated.

<sup>2</sup>Hayes-Roth, Frederick, Waterman, Donald A., and Lenat, Douglas B. (Eds.), *Building Expert Systems*, Addison-Wesley Publishing Company, Inc., London, 1983 p. 55.

<sup>3</sup>Nash, p. 361. A colleague had a similar experience. Their two and a half year old daughter had a much-loved doll. On a Friday night visit to a large store near closing time, the child saw a lone doll exactly like hers on a high shelf. The child wanted to get HER doll. My colleague ended up carrying her child screaming through the check-out, her daughter telling everyone in earshot that her doll would be lonely and cold over the weekend. On arrival at home, she rushed to her room, and returned with her adored doll clutched tight - "she beat us home, isn't she clever". The concept of "dolls in general" and "this doll in particular" had not yet been learnt.

<sup>4</sup>Biller, Henry and Meredith, Dennis *Father Power*, David McKay Company, Inc., New York, 1975, p. 225.

to conceive that the volume has not changed; the child is unable to reverse the thought process.

Some of the early classification expert systems had some of the characteristics of this latter limitation. Given the information that (e.g.) an animal had a long neck, long legs and a blotched coat, they could identify a giraffe. However, the process was irreversible — given that an animal was a giraffe, they could not provide information on the identifying characteristics.

Bee also comments that in the first two years of this range, the child (like some expert and control systems) only has *transductive* (specific to specific) reasoning power. With the child's reasoning:

Two things that happen together are taken to have some causal relationship. Piaget gives an example. Lucienne announced one afternoon when she had not taken her nap, "I haven't had my nap so it isn't afternoon." Afternoon and nap do usually go together, but she had the relationship between them wrong.<sup>1</sup>

This type of faulty reasoning can also occur when an adult examines the results of an expert system. As example of this type of faulty reasoning, consider an expert system related to (e.g.) pregnancy, which has been constructed from data collected as a result of a survey. In the survey mothers exhibiting early parturition may also check the boxes in the survey document relating to heavy smoking more frequently than others participating in the survey. This may result in the observation "early parturition occurred" being associated with the characteristic "heavy smoking" in an expert system built from this data. Although there is an adult tendency to assume an inductive/deductive chain of reasoning connecting the two items, (e.g. "heavy smoking" causes "early parturition")<sup>2</sup> the

---

<sup>1</sup>Bee, *The Developing Child*, p. 205.

<sup>2</sup>E.g. following Mill's "Rule of Concomitant Variations", see section 1.1.3, of this thesis. The reasoning involved is similar in principle to Russell's comment: 'If, whenever we can observe whether A and B are present or absent, we find that every case of B has an A as a causal antecedent, then it is probable that most B's have A's as causal antecedents, even in cases where observation does not enable us to know whether A is present or not.'; however in this case it has not been established that B has A as a causal antecedent, the relationship is only a transductive one. For the source of Russell's comment, see: Russell, Bertrand, 'Analogy', in Buford, Thomas O. (Ed.), *Essays on Other Minds*, University of Illinois Press, Urbana, U.S.A. 1970, p. 8.

relationship between the characteristic and the conclusion is purely transductive. One may be correlated with the other, but the relationship within the expert system, with smoking being one of the factors associated with premature labour, does not indicate cause and effect. There may well be cause and effect, but it must be proven elsewhere. The relationship within the expert system, like Lucienne's statement, is indicative only of reasoning at the transductive level.

If a child of this age is presented with two sticks with their ends level, the child can correctly identify them as being equal in length. Move one ahead of the other, and the child claims it is now longer, (perception taking priority over logic). Expert systems almost invariably have logic taking priority.

No ability to abstract has yet developed.<sup>1</sup> Abstract relationships present difficulties because children in this stage cannot grasp the hypothetical;

For instance, a six-year old child, when asked the question, "If your brother is a year older than you, how old is he?" protested that he could not answer this question because he did not have a brother.<sup>2</sup>

#### 1.2.5.3 *Stage of Concrete Thought.*

This occurs typically from seven to eleven years of age. At this stage the individual is capable of some logical operations where the logic is related to concrete instances. Note that logic of the type used by practicing logicians is almost invariably related to hypothetical situations, and this is beyond a person in the concrete stage. Within this fairly severe limitation, proper logical thought is possible. The individuals can also manipulate concepts if they are directly related to concrete reality, but are not yet able to deal with abstract propositions and hypothetical objects. The two stick problem mentioned above is a problem no more, as the evidence of logic can be accepted in preference to the evidence of perception. Individuals of this age can deal operationally and reversibly with the concept 'A is longer than B'. They can also

---

<sup>1</sup>Bee, *The Developing Child*, p. 205.

<sup>2</sup>Nash, p. 359.



deal with the concept of number, e.g. thirteen, involving the grouping of thirteen objects in a class and of relating or ordering the concept thirteen as being between the concept twelve and the concept fourteen. The individual also becomes capable of reversible thought — the clay problem can now be dealt with correctly.

An important additional ability is that the individual also now has the ability to think of all objects with a common feature together as forming a class of objects with that characteristic.<sup>1</sup> This is the ability that expert systems or classificatory systems attempt to simulate or emulate with inductive algorithms, and some decision tree-building algorithms may be compared with this stage.

#### 1.2.5.4 *Propositional or Formal Operations.*

This occurs from eleven years onwards to adulthood (although Piaget in his writing generally only refers to an upper age of 15 years). Deductive reasoning and hypotheses about hypothetical objects (rather than concrete ones) are now possible,<sup>2</sup> as Piaget states:-

The connection indicated by the words "if . . . then" (inferential implication) links a required logical consequence to an assertion whose truth is merely a possibility. This synthesis of deductive necessity and *possibility* characterises the use of possibility in formal thought, as opposed to possibility-as-an-extension-of-the-actual-situation in concrete thought ...<sup>3</sup>

Many expert systems usefully employ deductive logic operating near this level.

### 1.3 How widely applicable are these theories?

Inductive and deductive logic would not be of universal use to the community if these types of logic were only available to

---

<sup>1</sup>Inhelder, Barbel and Piaget, Jean *The Growth of Logical Thinking from childhood to adolescence*, Routledge & Kegan Paul, London, 1958, p. 105.

<sup>2</sup>Inhelder and Piaget, pps. 257-258.

<sup>3</sup>Piaget, Jean, *The Child's Conception of Physical Causality*, Kegan Paul, Trench, Trubner & Co. Ltd., London, 1930, p. 176.

individuals within the community selected by chance, inheritance or upbringing. In the following pages, section 1.3.1 poses the question as to whether everyone in the community develops deductive logic. Section 1.3.2 looks at the phenomenon of automaticity, and section 1.3.3 asks if inductive logic can help in cases of automaticity. This portion of the discussion then ends with section 1.3.4 noting the limitations of an imitative artificial intelligence or expert system even when it has the full power of transductive, inductive and deductive logic available to it.

### 1.3.1 Does everyone achieve deductive logic?

A word of caution is germane. Several authorities comment on the necessity of showing the chain of reasoning that leads to an expert system's conclusion, stating that an expert system will be unacceptable in practice without this feature. This chain will often be a series of 'if . . . then . . .' deductive logic statements, or a series of abstract rules. If expert systems are to be applied widely, it is worth noting that to understand this chain, it is necessary that the individual using the expert system has to have achieved this 'propositional or formal operations' stage. If not, they will not have the concepts to be able to deal with deductive logic. Bee comments -

Unlike the preceding stages, which seem to occur widely in many cultures, formal operations is achieved by only about half or two-thirds of the people in our culture, and by far fewer in less complex cultures.<sup>1</sup>

Cromer, a physicist with over three decades of University teaching, concurs. He comments:

... in a recent study of the mathematical ability of seventeen-year-olds in the United States, less than 6 percent could solve simple algebra problems (Saltus, 1989). It has been known for some time that most American college freshmen haven't reached the stage of formal operations (Lawson and Renner, 1974).<sup>2</sup>

---

<sup>1</sup>Bee, *The Developing Child*, pps. 222 - 223.

<sup>2</sup>Cromer, Alan, *Uncommon Sense*, Oxford University Press, Oxford, 1993, p. 26.

Many studies have shown that more than half of adult Americans never reach the stage of formal operations (Arons and Karplus, 1976), meaning they can't analyse a situation with several variables or understand a simple syllogism<sup>1</sup>

Holland *et. al.* refer to a review by Evans when noting that there is a controversy as to whether inferential rules are used by humans at all, noting that there is a:-

'body of evidence indicating that people are not able to make effective use of deductive rules of the kind that comprise the logic of the conditional when reasoning about abstract symbols'<sup>2</sup>

(In the terms of the previous discussion, these people could be roughly classified as being of Piaget's *concrete thought* stage.) However the authors argue against such a position, proposing:-

that people possess a wide variety of abstract, relatively domain-independent inferential rules that comprise pragmatic reasoning schemas. On the other hand we will argue that some extremely abstract inferential rules, notably those of formal logic, admit of so little application to real-world problems that people do not induce them and in fact cannot be easily taught to use them in pragmatic, everyday contexts.<sup>3</sup>

In terms of our previous discussion, Holland *et. al.*'s 'extremely abstract' category would appear to correspond to the higher end of the type of skills gained in Piaget's *formal operations* stage. Horizontal *décalage* suggests that people progress unevenly through this stage, it not being an 'all or nothing' development.

Let us consider those who do develop to the formal operations stage. Even with the resources implied by this stage, to gain full benefit from deductive rules given in an explanatory user interface, the user must also be cognisant with the special computer science meaning attached to the words 'if . . . then . . .', otherwise they may become confused about the meaning of the phrase. If Bree & Smit can list twelve distinct uses of the word

---

<sup>1</sup>*Ibid.*, p. 188.

<sup>2</sup>Holland, John H., Holyoak, Keith J., Nisbett, Richard E., Thagard, Paul R., *Induction*, The MIT Press, Cambridge, Massachusetts, 1987, pps. 44 - 45.

<sup>3</sup>*Ibid.*, p. 45, see also a fuller discussion pps. 255 - 286.

'if in English, there would reasonably seem a possibility of confusion amongst the non-computer-science literate.<sup>1</sup> The implications for an expert system requiring or using deductive logic, and aimed at a wide audience, are obvious.

There is also a problem in this area for computer scientists trying to implement artificially intelligent systems, particularly in the case of "common sense" systems. Suppose computer scientists operate predominantly at the propositional level (the lower levels which are also necessary for intelligence not being consciously accessible to them). If this were so, one would expect success in implementing activities similar to those undertaken by the computer scientists themselves, i.e. systems which, like the computer scientists, act as experts employing deductive logic. One would also expect less success in implementations which require the application of transductive and inductive intelligence (which were passed by the scientists on their way to the formal operations stage, and are now largely inaccessible) e.g. language, shape recognition, common sense. Alexander comments:

Adults often find the thought processes of young children incomprehensible because they [the adults] assume that formal and concrete operational assumptions are logically necessary and obvious. System designers find it relatively easy to produce formal operations in computers, but find it difficult to implement lower level human capabilities such as language on such systems. Perhaps this is because they have set themselves the fascinating challenge of trying to build intelligence backwards without thinking about what it is.<sup>2</sup>

In relation to attempts to build "common sense" systems, it is worth noting Bee's comments that over one third to one-half of people in our culture do not achieve the propositional stage, and hence the concomitant availability of abstract deductive logic. Unless one is prepared to suggest that up to half the population

---

<sup>1</sup>Bree, D. S & Smit, R., *Non Standard Uses of IF*, in Elithorn, Alick and Banerji, Ranan (Eds.), *Artificial and Human Intelligence*, Elsevier Science Publications B. V., Amsterdam, 1984, pps. 317-318.

<sup>2</sup>Alexander, James, *Intelligence: Natural and Artificial*, Seminar handout, Hobart, Tasmania, 15 October 1992, p. 10.

do not exhibit common sense,<sup>1</sup> it would seem likely that in humans common sense results from the application of transductive or inductive logic. If Alexander is correct, the implications of this for computer scientists attempting to build a "common sense" knowledge base employing deductive logic may well be profound.<sup>2</sup>

### 1.3.2 Automaticity

Automaticity can be a problem for knowledge engineers who are attempting to amass a collection of rules which describes an expert's knowledge.

If these rules are part of a Gestalt, the expert may not be conscious of them as individual entities, may not have to think about them, he may just automatically apply them. 'An expert is one who does not have to think. He knows.'<sup>3</sup> Similarly, Pine comments 'The pinnacle of expertise in a field is intuition in that area.'<sup>4</sup>

Alternately, the knowledge may be transductively or inductively derived from the expert's experiential knowledge base, in which case abstract deductive rules of the type sought by the knowledge engineer building an expert system may not be either available from the expert, or necessary to allow the expert to function as an expert.

Computer scientists have talked about this type of knowledge being "compiled knowledge", the implication being that while the results of the original instructions (program) is available for use, the instructions themselves are not. In this case they are implicitly assuming that the rules existed at one time in the expert's mind, because (following Alexander) that is the way the computer scientists would generally think about it themselves.

---

<sup>1</sup>Views stronger than this have been held by significant philosophers, e.g. Blakemore comments 'Plato was a wealthy aristocrat who believed leisure was essential to wisdom, which was therefore automatically denied to the working poor.', see: Blakemore, Colin, *Mechanisms of the Mind*, Cambridge University Press, Cambridge, 1977, p. 12.

<sup>2</sup>It is interesting to compare this idea of common sense with the Dreyfuses' gestalt-like concept of human information processing; see section 1.2.4 of this thesis.

<sup>3</sup>Wright, Frank Lloyd, quoted in Minsky, Marvin, *The Society of Mind*, Simon and Schuster, New York, 1986, p. 137.

<sup>4</sup>Pine, Milton, 'Western Philosophy and Expert Systems', *Professional Computing*, Peter Isaacson Publications, Victoria, Australia, October 1989, p. 27.

While this is possible, it is also possible that the rules never existed in the expert's mind in this form at all, (or needed to).

R. B. Cattell also noted the existence of this type of knowledge. He developed a theory of intelligence which included both fluid and crystallised intelligence, the latter addressing phenomena similar to automaticity. Alexander comments:

Fluid intelligence ... is the fundamental capability to induce relationships, fluid in the sense of being able to be directed to almost any intellectual problem, but best measured by tests of inductive reasoning. ... Crystallised intelligence ... is the product of experience, learned factual content and problem solving strategies, ... According to R. B. Cattell, crystallised intelligence develops through the investment of fluid intelligence in the acquisition of skills and knowledge subject to environmental opportunity.<sup>1</sup>

To illustrate the differences between the application of these types of abilities, let us artificially divide experts into two groups, those who "know", and those who "do".

An expert at a University can usually explain the basis for his reasoning, being an academic expert skilled in the art of verbal expression.

An expert at riding a bicycle may have a problem in explaining exactly how (s)he rides that bicycle, being an expert at doing. Automaticity is more a factor in the latter case than in the former.

Mishkin and Appenzeller discuss the phenomenon of automaticity.<sup>2</sup> They postulate a second system of learning, independent of the limbic circuits [which would appear to be the main system involved in Piagetian development]. This neo-Pavlovian learning, sometimes referred to as automaticity, is of the repetitive stimulus-response type, probably mediated by the (in evolutionary terms ancient) striatum, although Groves and Schlesinger cite experimental evidence that demonstrates that

---

<sup>1</sup>Alexander, James, p. 6.

<sup>2</sup>Mishkin, Mortimer, & Appenzeller, Tim *The Anatomy of Memory*: Scientific American, Scientific American Inc., June 1987, Vol. 256, No. 6, p.71.

mediation by the spinal cells alone may be sufficient in some cases.<sup>1</sup> This type of learning:-

is non-cognitive; it is founded not on knowledge or even on memories (in the sense of independent mental entities) but on automatic connections between a stimulus and a response.<sup>2</sup>

Although being 'gifted' as a child does not necessarily lead to being a 'gifted' adult,<sup>3</sup> (and hence possibly an expert), it is interesting to note it has been 'hypothesized that gifted children are superior in ... automatization'.<sup>4</sup>

If this is so, it is probably important for the knowledge engineer attempting to get an expert to express his or her expertise in the form of deductive rules to know that :-

if neural mechanisms for both kinds of learning do exist, behaviour could be a blend of automatic responses to stimuli and actions guided by knowledge and expectation.<sup>5</sup>

This blend may be one of the expert's strengths, in that it allows a rapid response. However the concomitant limitation which this may imply is the expert's inability to access this information stored in a non-consciously-accessible form.<sup>6</sup>

---

<sup>1</sup>Groves, Phillip, and Schlesinger, Kurt, *Biological Psychology*, Wm. C. Brown Company, Dubuque, Iowa, 1979, p. 469.

<sup>2</sup>op. cit..

<sup>3</sup>The correlation, however, is high. Mike Anderson comments (p.7) "that IQ measured at 5 years old predicts around 50 per cent of the variance in mathematics scores at 16 ... The year-to-year correlation between IQ scores is remarkably high, around 0.9, while over the whole period of schooling it is approximately 0.7"

<sup>4</sup>Siegler, R. S., & Kotovsky, K., 'Two levels of giftedness' in Robert J. Sternberg & Janet E. Davidson (Eds.), *Conceptions of Giftedness*, Cambridge University Press, Cambridge, 1986, p. 422.

<sup>5</sup>Groves & Schlesinger, p. 469.

<sup>6</sup>This has been commented on in widely separated fields, e.g. to give three examples:-

1) In the medical field:-

Dr. William Mouradian, when talking about his efforts to express diagnosis in terms of rules, comments 'As an intern, I was disenchanted by the inability of many of my instructors to explain their decisions', but when later, as an instructor, attempting to do so himself he 'had enormous difficulty breaking down my thought processes into rules. With considerable effort, I verbalised the most general rules used in my decision making. Clinicians use many rules of inference, but they will have difficulty enunciating them to a knowledge engineer' because 'many of the observations are of non-verbal behaviour' and 'the clinician will have difficulty articulating the details of their observations and reasoning, remembering only the overall impression. Decision making in medicine is quite intuitive'; from Mouradian, William H., 'Knowledge

Hofstadter appears to not only accept a separation of the logical and thinking levels, but suggests the separation may be advantageous:-

Luckily for you, your symbol level (i.e. *you*) can't gain access to the neurons which are doing your thinking—otherwise you'd get addle-brained. To paraphrase Descartes again:

"I think; therefore I have no access to  
the level where I sum"<sup>1</sup>

The expert's 'insights... are private and, except through symbols and at second hand, incommunicable'.<sup>2</sup> If the symbols are unattainable, the insights are incommunicable.

---

Acquisition in a Medical Domain', *AI EXPERT*, Vol. 5, No. 7, July 1990, pps. 36-37.

2) In the field of human-machine interfaces:-

Mayes et. al. comment on a similar phenomena which was "exemplified by the observation that some skilled touch-typists are not aware at a conscious and reportable level of the layout of the keyboard: if asked where, say, 'X' is located they have to imagine it and follow the finger movement. Since they must have observed the finger movement by using visual search of the keyboard, we might interpret this as a 'compiling-in' of action sequences and the dropping (eventual forgetting) of the visual representation they were derived from. That is, the visual representation *became* 'incidental' and hence is forgotten, even though it was once a necessary part of performance" (p. 231). They also report a similar phenomena with relation to use and recall of items in the menus of the Macintosh interface (p. 230). They further comment on the possibility that the recall may not even be available under all circumstances, as these could represent "an example of encoding specificity (Tulving, 1974) where recall is only possible when the retrieval cues present at learning (at the interface) were also present at recall" (p.232); from Mayes, J. Terry, Draper, Stephen W., McGregor, Alison M. and Oatley, Keith, "Information Flow in a User Interface: The Effect of Experience and Context on the Recall of MacWrite Screens", in Preece, Jenny and Keller, Laurie (Eds.), *Human-Computer Interaction*, Prentice Hall, Hertfordshire, England, 1989. For completeness, the Tulving reference is included in the reference list of this thesis.

3) In the Industrial area:-

Vaux comments on the experience of a professional photographer Peter Gullers, who adjusts his camera "without the benefit of an automatic light meter. The judgement he makes is based on years of experience, but 'all of these earlier memories and experiences that are stored away over the years only partly penetrate my consciousness . . . The thumb and index finger of my right hand turn the camera's exposure knob to a setting that 'feels right', while my left hand adjusts the filter ring. This process is almost automatic'. The rules he follows are expressed directly in action; they are not a set of propositions, not even a set of formulae for calculating the f-stop." (pps. 40-41). Vaux refers to this (and skills such as walking across a room, and driving a car) as "examples of implicit knowledge: . . . tasks that humans find easier to do than to describe" (p. 40); from Vaux, Janet, "Replicating the expert", *New Scientist*, 3<sup>rd</sup> March 1990, pps 39-42.

<sup>1</sup> Hofstadter, Douglas R., *GÖDEL, ESCHER, BACH: An Eternal Golden Braid*, Penguin Books, Harmondsworth, England, 1982, p. 677.

<sup>2</sup> Huxley, Aldous, *The Doors of Perception*, in *The Doors of Perception and Heaven and Hell*, Penguin Books, Harmondsworth, England, 1961, p. 13.



This is important, as in this case of automaticity the problem goes deeper than Partridge's suggestion of a different internal symbolic representation to that used in a production rule system that would lead to an inability to express expertise in production rule format.<sup>1</sup> It would suggest that in some cases there is *no* symbolic independent mental representation of the learning, and hence that, in this case, the expert would be incapable of expressing his or her expertise in either a production rule or any other theoretical format.<sup>2</sup>

### 1.3.3 Automaticity and Induction

The implications are obvious concerning a knowledge engineer who wants the expert to formulate rules in the deductive format so convenient for the knowledge engineer, and may be the reason Modesitt comments 'It is terribly difficult for an expert to explicate her/his knowledge in a rule format'.<sup>3</sup> In some cases this information may just not be available, because of limitations concomitant with the expert's strengths; Modesitt continues 'the more expertise they have, the more difficult it is for them to articulate their knowledge'.<sup>4</sup> Gevarter observes 'Human experts are often able to articulate their expertise in the form of examples better than they are able to express it in the form of rules'.<sup>5</sup>

---

<sup>1</sup>Partridge, p. 349.

<sup>2</sup>At this stage, one might reasonably ask, "Why are rules necessary?". The answer, in practice, seems to be two-fold. Firstly, computer scientists need them. In the author's experience academic computer scientists from a maths/physics background operate almost exclusively in a propositional mode. Hence they need deductively-expressed rules to understand; inductively-derived rules do not constitute understanding to someone who operates only at the propositional level. Hence the need may be the scientist's need, rather than the system's need. The second reason is that (compared to the human brain) the very small storage spaces available in practical computers make the space and speed saving ability of deductively expressed rules vital in practice. A single rule may summarise an aspect of a huge amount of experiential data, and a computer can generally activate a rule much more quickly than it can search a large data base. By contrast, humans have relatively slow computing elements, but (by computer terms) remarkable pattern-matching abilities; ideal for assisting transductive and inductive logic.

<sup>3</sup>Modesitt, K. L., 'Experts: Human and Otherwise', *Proceedings of the Third International Expert Systems Conference*, Learned Information Ltd., Oxford, 1987.

<sup>4</sup>Op. Cit.. There is an interesting parallel here with neural networks. In neural networks the knowledge is expressed in a series of dendritic weights, not in the sequential if...then type of rules employed by many expert systems. Some authorities regard this knowledge as inaccessible because it is not expressed in logical rules, and have attempted (with limited success) to extract "real knowledge" (rules) from the neural net weights.

<sup>5</sup>Gevarter, William B., 'The Nature and Evaluation of Commercial Expert System Building Tools', *Computer*, Volume 20, Number 5, I.E.E.E., New York, May 1987, p. 27.

In these cases, an *inductive* process based on instances of the expert's actions may well be appropriate, as it can (partially) transcend the expert's limitations. The problem caused by these limitations has become known as the Feigenbaum bottle-neck.<sup>1</sup>

### 1.3.4 Limitations of Expert Systems

But while the experts may be limited, so are the expert systems. There is still a wide gap between an artificial intelligence system employing transductive, inductive and deductive logic and Smart and Smart's characterisation of an adolescent in the propositional or formal logic stage, who,

instead of having to base his thoughts on concrete things and events, he is thus freed from restraints of time and space, able to range throughout the universe, entertaining concepts with which he has had no real experience, such as the notion of infinity . . . he does not get stuck with his perceptions as does the preschool child, or stuck with his conclusions, as does the school-age child . . .<sup>2</sup>

While artificial intelligence cannot yet simulate the full powers of the adolescent, it can do a reasonable to good job in specific areas.

## 1.4 Simulation of Induction

Now that some background to the theory of induction has been given, it is appropriate to examine briefly how a human uses induction (section 1.4.1), and how uses of those inductive powers can be tested (section 1.4.2). In this latter section, examples of the testing of inductive logic in IQ (intelligence quotient) tests are given. Finally the advantages that the use of induction bring to human problem solving are noted in Section 1.4.3.

### 1.4.1 What can be simulated?

In Piaget's genetic epistemology, the concept of induction is developed by the individual after transductive logic and usually before deductive logic, (the spread of horizontal décalage

---

<sup>1</sup>Quinlan, *Induction of Decision Trees*, p. 2.

<sup>2</sup>Smart, Mollie S. & Smart, Russell C., *Children: Development & Relationships*, Macmillan Publishing Co., Inc., New York, 1977, p. 526.

between individuals may produce some exceptions to the latter statement where individual abilities are concerned, but usually this statement is accurate).

Pellegrino defined induction as 'the development of general rules, ideas or concepts from sets of specific instances or examples'.<sup>1</sup> Human children (and some animals) obtain these sets by initially using trial and error methods, later forming a strategy. Harlow comments that this is 'learning to learn' and 'learning to think'.<sup>2</sup> We learn:-

by analyzing the similarities and differences between specific experiences, we extract the general characteristics of classes of objects, events and situations. We apply these generalizations to new experiences, refine and modify them, and make them part of our permanent knowledge base.<sup>3</sup>

In Piagetian terms, the new information would be assimilated into the person's schema,<sup>4</sup> categorisation permitting the new experience to be related to other portions of the schema so the experience became part of a consistent whole, the age and speed at which this occurs depending in part on the individual's horizontal décalage.<sup>5</sup>

#### 1.4.2 Relationship between Induction and Intelligence

It will be assumed in this section that intelligence and IQ (Intelligence Quotient) are closely correlated. Since this is a controversial assumption, some explanation should be given.

Alfred Binet developed tests based on the concept of mental age in 1904. These tests were the pre-cursors of what were to

---

<sup>1</sup>Pellegrino, James W., *Inductive Reasoning Ability*, in Sternberg, Robert J. (Ed.), *Human Abilities*, W. H. Freeman & Company, New York, 1985, p. 195.

<sup>2</sup>Harlow, Harry F. and Harlow, Margaret Kuenne, *Learning to Think*, in *Physiological Psychology, Readings from the Scientific American*, W. H. Freeman and Company, San Francisco, 1972, pps. 401 - 405.

<sup>3</sup>loc. cit..

<sup>4</sup>Either slowly and reversibly, through inductive integration, or rapidly and irreversibly, through noetic integration, Beth, Evert W. & Piaget, Jean, *Mathematical Epistemology and Psychology*, D. Reidel Publishing Company, Dordrecht, Holland, 1966, p. 126.

<sup>5</sup>Brainerd, Charles J., *Piaget's Theory of Intelligence*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978, p. 36, also Labinowicz, Ed, *The Piaget Primer, Thinking, Learning, Teaching*, Addison-Wesley Publishing Company, Menlo Park, California, 1980, p. 92.

become intelligence tests. If one of his tests distinguished between individuals he thought were of different mental age, he retained it, if it didn't, he discarded it. This was a "practical" as opposed to a "theoretically acceptable" approach. Another example of a "practical" approach is Dolbear's law. If one counts the number of chirps a snowy tree cricket produces in 15 seconds, and adds 40, one has the temperature in degrees Fahrenheit. Whilst this is a practical tool if one has a supply of snowy tree crickets, it is much easier for a present-day physicist to theoretically justify a thermometer based on the expansion of the liquid metal mercury.<sup>1</sup> The arguments about IQ tests basically centre on whether they are nearer "Dolbear's law" or "thermometers" in their theoretical validity, and exactly what they measure.

In this thesis we will accept a Dolbear's law approach to IQ tests; that is, regardless of exactly how they work, they produce a practically useful result which we will accept as an indication of human intelligence for the purposes of this discussion.

If one considers the relationship between induction and human intelligence, it is fair to state that while not all 'learning by experience' in computers or humans involves induction, (e.g. this author's unpublished work simulating Pavlovian conditioning in 1965), the concept of inductive reasoning ability has historically been central to theories of human intelligence. Indeed Pellegrino notes that Thurstone considered induction to be one of the primary mental abilities.<sup>2</sup> It is used widely in IQ and ability tests. Pellegrino comments that 'one or more of these tasks can be found on virtually any current aptitude or intelligence test at any age level . . .'.<sup>3</sup>

As an example, in the Cognitive Abilities Test (CAT) test aimed at grades three to twelve, inductive ability subtests constitute 50% of the entire Test, testing induction in the areas of verbal analogy, verbal classification, figural analogy, figural classification, and number series. Examples of the types of

---

<sup>1</sup>To be fair, the thermometer was also originally developed as a "practical" tool, but this is irrelevant to the present-day situation posited in the above argument.

<sup>2</sup>ibid., p. 196.

<sup>3</sup>loc. cit.. Note that I.Q. tests are not generally used with children whose mental age is below about 5 years.

problems are shown in Figure 1.<sup>1</sup> The proportion of inductive subtests is important, because, whatever psychological theory states, the operational definition of intelligence used by practising psychologists would seem suspiciously close to Boring's terse comment that 'Intelligence is what intelligence tests measure'.<sup>2</sup>

It will be noted that analogical reasoning is one of the subtypes of inductive reasoning, and Sternberg notes Raven designed his Progressive Matrices Test as a largely culture-fair<sup>3</sup> "test . . . [of] . . . a person's present capacity to form comparisons, reason by analogy, and develop a logical method of thinking".<sup>4</sup> This test uses figural analogies of the type shown in Figure 2.<sup>5</sup>

---

<sup>1</sup>Pellegrino, p. 197.

<sup>2</sup>Quoted by Sternberg, Robert J., *General Intellectual Ability*, in Sternberg (Ed.), p. 21.

<sup>3</sup>English & English, p. 547, define a culture-fair test as "a test of general ability from which have been eliminated, as far as possible, all items depending on experiences that are more commonly found in one culture than another. Such tests must eliminate language, and the information or skills selectively employed in one culture more than in others".

<sup>4</sup>Pellegrino, p. 199.

<sup>5</sup>*ibid.*, p. 204.

Figure 1 — Examples of inductive reasoning problems

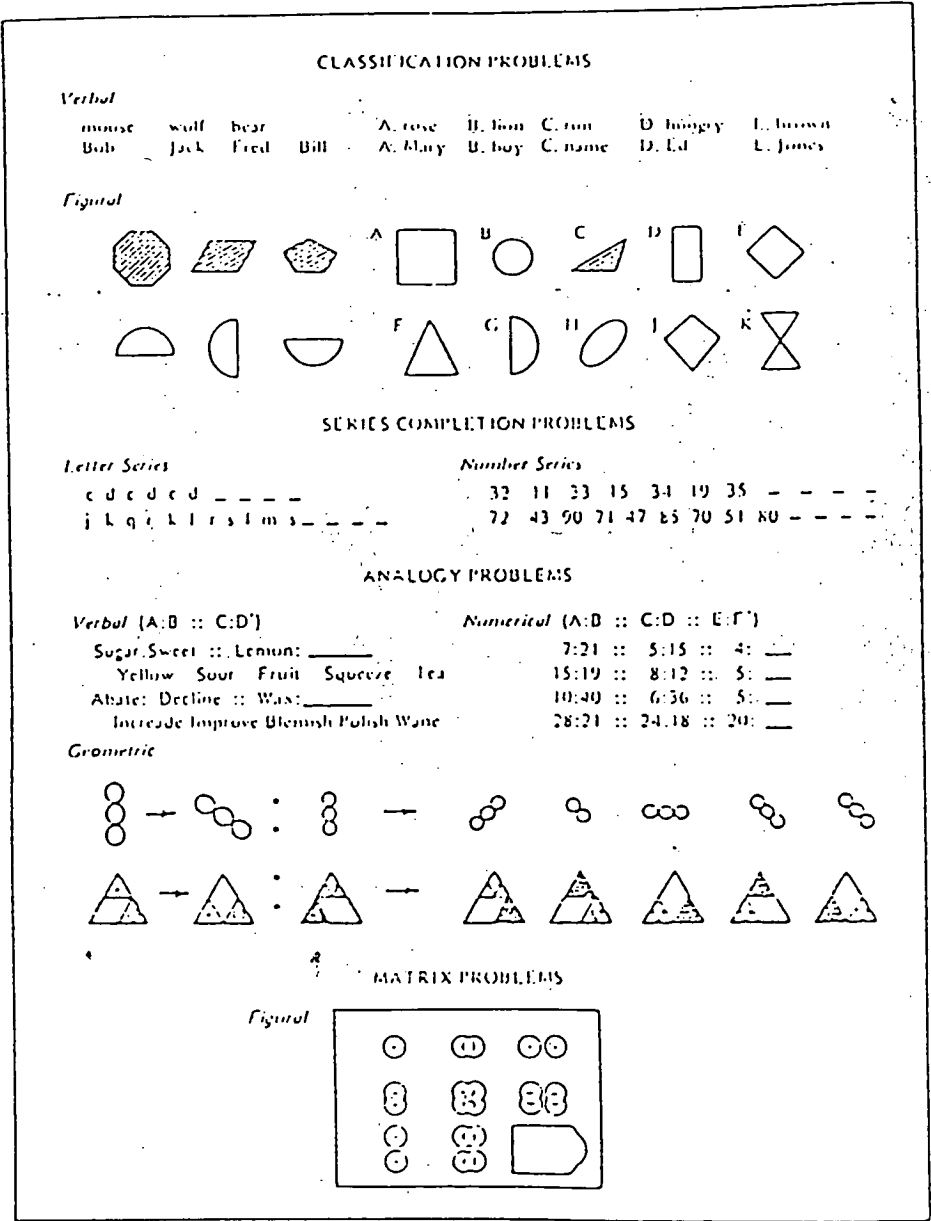
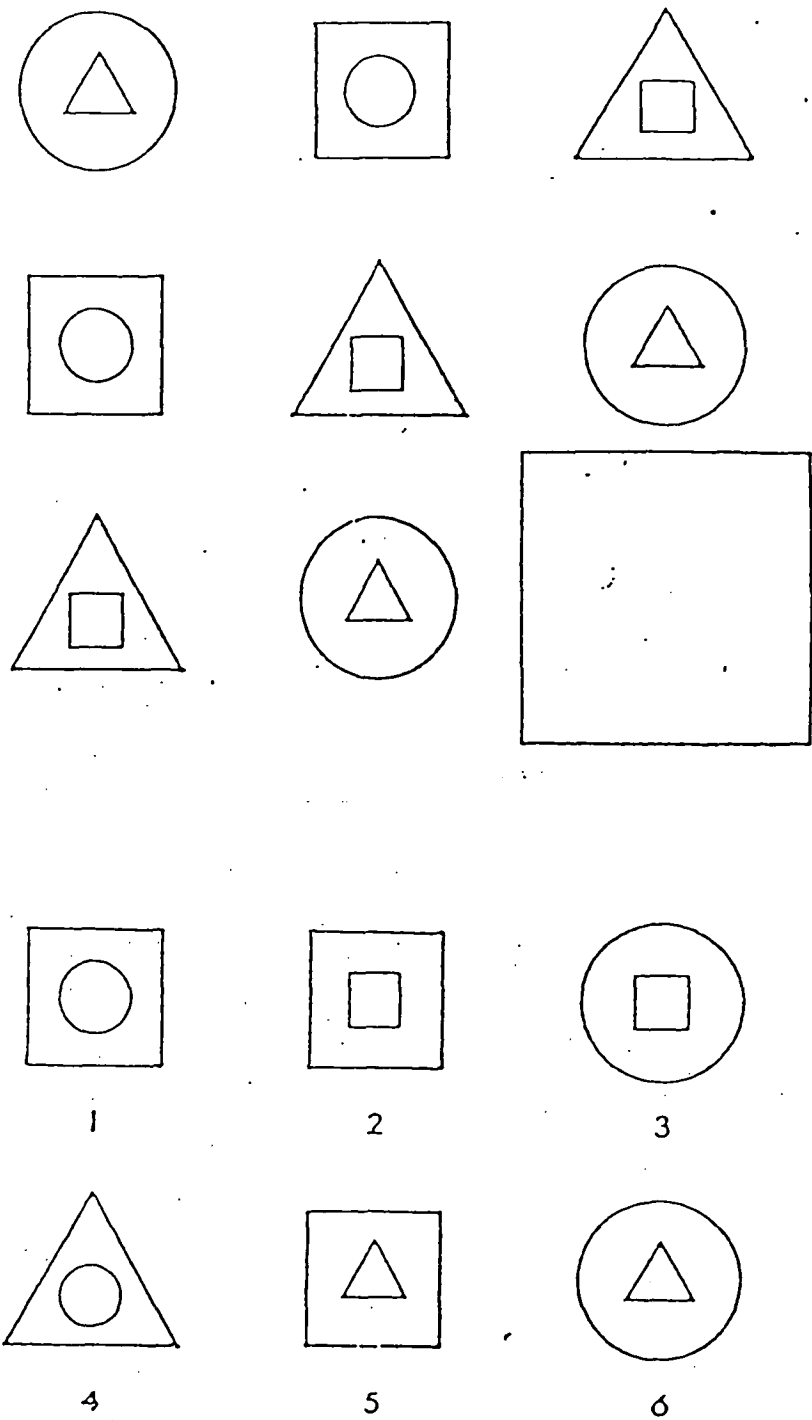


Figure 2 — Examples of true and false figural analogies

Item class	True analogies	False analogies
1 Element 1 Transformation		
1 Element 3 Transformations		
2 Elements 2 Transformations		
3 Elements 1 Transformation		
3 Elements 3 Transformations		

Figure 3 — An IQ test analogical reasoning problem

1. Which of the six numbered figures fits into the vacant square? (Insert the number in the square.)





Winston<sup>1</sup> cites work by Evans on analogical reasoning, and describes a procedure which, when implemented on a computer, effectively simulates human inductive analogical reasoning when handling figures of the type shown in Figure 3.<sup>2</sup> Apparently this work permits simulation of human analogical reasoning to approximately school leaving ability.

To solve this type of analogical reasoning problem, the subject must first inductively categorise or classify the given examples, and then attempt to apply the categorisation criteria to the other examples in an attempt to include one of them (see Figure 1, verbal and figural classification, verbal and geometric analogies) or to inductively or deductively extend the series (Figure 1, Series Completion, numerical analogy and figural matrix, or Figure 3).

#### 1.4.3 Advantages of Induction

Watson and Lindgren comment that the cognitive advantage gained as the result of conceptual inductive categorisation can be summarised in five points:-<sup>3</sup>

- a) the complexity of the environment is reduced;
- b) provision is made for a means of identification of objects in the environment;
- c) the necessity of relearning at each new encounter is reduced;
- d) help for the direction, prediction and planning of any activity is provided; and
- e) ordering and relating classes of objects and events as in cause and effect is provided.

---

<sup>1</sup>Winston, Patrick Henry, *Artificial Intelligence*, (edn. 1), Addison-Wesley Publishing Company, Reading, Mass., 1979, pps. 16 - 28; Winston, Patrick Henry, *Artificial Intelligence*, (edn. 2), Addison-Wesley Publishing Company, Reading, Mass., 1984, pps. 24 - 35.

<sup>2</sup>Eysenck, H. J., *Know Your Own IQ*, Penguin Books, Harmondsworth, Middlesex, 1962, p. 118; see similar examples in Butler, Eamonn and Pirie, Madsen, *Test Your IQ*, Pan Books, London, 1983.

<sup>3</sup>Watson & Lindgren, p. 269.

In short, conceptual inductive categorisation is a prerequisite for deductive reasoning in humans. It is worth re-emphasising however that induction can be a powerful tool in its own right, many operations being able to be handled quite concretely,<sup>1</sup> admittedly with *concrete* persons adopting less complex stratagems than *formal* people.<sup>2</sup> Sloman even prefers inductive to deductive logic in some circumstances, commenting,

It seems that the human brain is made from relatively slow computational units, although there are very many of them. This means that if recognising dangerous situations, or working out what to do, requires long chains of reasoning from general principles, then, before decisions are taken, death or other disasters may ensue.<sup>3</sup>

He goes on to suggest that the results of reasoning by oneself or others may be stored, ready for 'blind' recall and immediate use. This process seems akin to an inductive classification of the situational characteristics, together with the appropriate response. The prompt response provided by a mechanism like this would give its possessor a marked advantage in time-critical and/or dangerous situations, and could be expected to be strongly favoured from an evolutionary stand-point. It is thus of little surprise that inductive logic is developed preferentially to deductive logic in almost all humans (some of whom subsequently go on to also develop a deductive facility). Some other writers differentiate between an expert's 'intuitive' and 'logical' knowledge. It is possible that inductive mechanisms of the type described above may be behind the so-called 'intuitive' logic.<sup>4</sup>

---

<sup>1</sup>Vinacke, W. Edgar, *Foundations of Psychology*, American Book Company, New York, 1986, p 391.

<sup>2</sup>Vinacke, p. 609.

<sup>3</sup>Sloman, *Towards a computational theory of mind*, in Yazdani, Masoud and Narayanan, Ajit (Eds.), *Artificial Intelligence: human effects*, Ellis Horwood Limited, Chichester, 1984, pps. 173-181.

<sup>4</sup>If this is so, the knowledge engineer's desire to obtain from an expert a rule expressed in abstract deductive format may be doomed to failure if the expertise held by the expert is held in inductive form.

## 1.5 Controversy about existence of Artificial Intelligence

Having discussed some of the aspects of intelligence in humans, the next question would be to ask if something like this human intelligence can be simulated or expressed in or from a machine. Opinions about this possibility vary widely, from strongly positive to strongly negative. In this section opinions about the possibility will be examined, as well as some axiomatic beliefs which, it is argued, may well provide the observer with an (often unexamined) pre-existing bias on this question.

Simon comments that the term *artificial intelligence* was coined at M.I.T..<sup>1</sup> Anderson claims the term was invented 28 years ago by Professor John McCarthy, now a computer scientist at Stanford.<sup>2</sup> Hilts concurs.<sup>3</sup> Despite the age of the concept, there is still some disagreement as to what constitutes artificial intelligence, and even the idea seems to provoke unease.

... most people have serious misgivings about the feasibility and, more importantly, the desirability of attributing the actions of a machine to intelligent behaviour. These people generally distrust the concept of machines that approach (and thus why not pass?) our own human intelligence. In our culture an intelligent machine is immediately assumed to be a bad machine.<sup>4</sup>

That the concept of intelligence in machines should provoke misgivings, and that something as seemingly morally neutral as a machine should be classified as bad, deserves examination. The type of classification accorded intelligent machines may be predicated by the (often unexamined) philosophical axioms of

---

<sup>1</sup>Simon, Herbert A., *The Sciences of the Artificial*, The MIT Press, Cambridge, Mass., 1985, footnote p. 6.

<sup>2</sup>Anderson, Ian, "AI is start naked from the ankles up", *New Scientist*, 15 November 1984, IPC Magazines Ltd., England, 1984.

<sup>3</sup>Hilts, Philip J., 'The Dean of Artificial Intelligence', *Psychology Today*, Volume 17, number 1, January 1983, p. 28.

<sup>4</sup>Negroponte, Nicholas, quoted by Baecker, R.M., Buxton, W. A. S., "An Historical and Intellectual Perspective", in Preece, Jenny and Keller, Laurie (Eds.), *Human-Computer Interaction*, Prentice Hall, Hertfordshire, Great Britain, 1990, p. 19. Baecker and Buxton do not give the source of their quotation, but comment that an elaboration of Negroponte's views can be found in Negroponte, Nicholas, *Soft Architecture Machines*, Cambridge, MA., the MIT Press, 1975.

belief employed by the person making the judgement. It will be argued that some sets of axiomatic beliefs make the idea of machine intelligence quite unacceptable, while others put no obstacle in the path of the acceptance of the concept of intelligent machines. Since this argument is not a main plank of this thesis (and is presented as part of the background of this thesis) it will be argued more briefly than might otherwise be justified; more detail may be found by consulting the texts referenced in the footnotes.

### 1.5.2 Beliefs taken as axioms, and their consequences.

Some philosophers claim 'that there is no absolute or objective truth ... something is true if we believe it to be true'.<sup>1</sup> They divide people who take this view into two categories, relativists and nihilists. Nihilists assert there is no truth at all. Relativists assert that truth is relative, e.g. an anthropologist may accept the beliefs of a foreign culture as being "truth for them". For the purpose of this thesis, the author rejects the nihilist view,<sup>2</sup> and accepts the relativist view, but with the codicil that for the person holding those beliefs their beliefs can well be 'beliefs [that] are truth, not just "truth for them"'.<sup>3</sup> Chwedorowicz comments on how these 'truths' or 'axioms' are internalised:

Beliefs are never isolated in a person's space of knowledge, they rather combine and come together as systems. It is characteristic of these systems that they have mechanisms for protecting the cohesion of each system. ... All the mechanisms may be described logically in the forms of axioms of a system of beliefs. The axioms are descriptions of certain rules according to which the real system of belief works. Accordingly, as these

---

<sup>1</sup>Heathcote, Adrian, "False prophets muddy 'truth'", *The Australian*, 9 March 1994, Nationwide News Pty. Ltd., Canberra, 1994, p. 29. See also Cromer, who comments 'Academics cringe at the words *truth* and *certainly*. They believe that truth and certainty aren't possible because philosophers have shown that neither empirical nor deductive knowledge can be made error free.' Cromer, p. 17. (The italics were in the original source).

<sup>2</sup>More detailed discussions of the concept of truth can be found in many philosophical textbooks, but as an example of an argument against the nihilist belief, consider the Tasmanian judicial system which is based on the concept of 'truth'. If a person takes an oath or affirms to *tell the truth, the whole truth and nothing but the truth*, it at the very least implies that such a truth exists. To deny the existence of truth would be to go against common sense in that it would place the judicial system in jeopardy.

<sup>3</sup>Heathcote, p. 29.

rules are fulfilled, information is accepted and interiorized, or rejected, or may make the system change'<sup>1</sup>

In this thesis Chwedorowicz's assertion that these "truths" or "rules" form a set of axiomatic beliefs that underpin a person's belief system, is accepted. It is also noted that such a set of axiomatic beliefs implicitly involves an ontological commitment about the nature of the world.

Let us consider a sub-set of such a set. If *determinism* is the belief that every event has a cause, consider the following three propositions (truths, rules, axioms, beliefs?):

1. Determinism is true.
2. If determinism is true, then no actions are free.
3. Some actions are free.<sup>2</sup>

A consequent of determinism is that every event is, in principle, predictable. If every event is predictable, then proposition 3. above is not true, and instead a person accepting 1. and 2. would substitute "No actions are free"<sup>3</sup> for proposition 3.. People accepting propositions 1. and 2., but rejecting 3. are often called *hard determinists*. People accepting propositions 1. and 3., but rejecting 2. are called *soft determinists* or *compatibilists*.<sup>4</sup> People accepting propositions 2. and 3., but rejecting 1. are called *libertarians*.<sup>5</sup>

---

<sup>1</sup>Chwedorowicz, Józef, 'Origin, structure and function of fuzzy beliefs', in Zétényi, Tamás (Ed.), *Fuzzy Sets in Psychology*, North-Holland, Amsterdam, 1988, p. 276.

<sup>2</sup>Klemke, E.D., Kline, A. David, and Hollinger, Robert (Eds.), *Philosophy The Basic Issues*, St. Martin's Press, New York, 1982, p. 100. Also see a related statement by St Augustine in: Augustine, St., 'The Freedom of the Will', in Berofsky, Bernard (Ed.), *Free Will and Determinism*, Harper and Row, New York, 1966, pps. 271-272.

<sup>3</sup>Ibid.

<sup>4</sup>Discussing soft determinism, Taylor re-states proposition 3. as "voluntary behaviour is nonetheless free to the extent that it is not externally constrained or impeded"; see Taylor, Richard, "Freedom and Determinism", in Klemke, et. al., pps. 118, 119.

<sup>5</sup>These concepts can only be discussed briefly here. For further discussion see texts such as Dennett's very readable book on 'The Varieties of Free Will Worth Wanting', Dennett, Daniel C., *Elbow Room*, The MIT Press, Cambridge Massachusetts, 1984; Berofsky, Bernard (Ed.), *Free Will and Determinism*, Harper and Row, New York, 1966; and Trusted, Jennifer, *Free Will and Responsibility*, Oxford University Press, 1984.

It will be argued that persons holding views which would classify them as determinists would find less difficulty accepting the concept of machine intelligence than would libertarians; and that some libertarians would reject such a concept immediately as violating a theorem consequent of the beliefs they take as axiomatic.

Major problems for the concept of machine intelligence include the *algorithmic objection*, and arguments grouped under the headings of *qualia* and *intentionality*.

The "quale" of a mental state or event is that state or event's *feel*, its introspective "phenomenal character." Many philosophers have objected that neither . . . . AI nor the computer model of the mind can explain, illuminate, acknowledge or even tolerate the notion of *what it feels like* to be in a mental state of such-and-such a sort.<sup>1</sup>

"Intentionality" is a feature common to most mental states and events, particularly the "propositional attitudes," ... propositional attitudes *represent* actual or possible states of affairs. ... One believes *that broccoli is lethal*, desires *that visitors should wipe their feet*, hopes *that the Republican candidate will win*, etc. Other propositional attitudes include thoughts, intentions, rememberings, doubts, wishes and wonderings.<sup>2</sup>

Only the algorithmic objection will be considered here, primarily because it is (in the author's experience) the one most frequently raised amongst computer scientists. Consideration of the other objections in any detail would need a thesis in itself.<sup>3</sup>

---

<sup>1</sup>Lycan, William G. (Ed.), *Mind and Cognition*, Blackwell, 1990, p. 10. Qualia are discussed in many philosophical texts, e.g. see a brief treatment in: Putnam, H., 'Robots: Machines or artificially created life', in Crosson, Frederick J., (Ed.), *Human and Artificial Intelligence*, Appleton-Century-Croft, New York, 1970, pps. 177-202.

<sup>2</sup>op. cit..

<sup>3</sup>See discussions on these topics in books such as Hofstadter, Douglas R. and Dennett, Daniel C., (Eds.), *THE MIND'S I*, Penguin Books, Harmondsworth, England, 1982; and Crosson, Frederick J., *Human and Artificial Intelligence*, Appleton-Century-Crofts, New York, 1970.

The algorithmic objection states that the computer is just following an algorithm and hence intelligence cannot be involved. Harel concurs, commenting:

We tend to view intelligence as our quintessential nonprogrammable, and hence non-algorithmic, feature.<sup>1</sup>

In other words, if behaviour can be reduced to an algorithm, (the results hence being completely predictable) intelligence simply cannot be involved. It is thus argued that the computer is simply an automaton, and that automatons cannot exhibit intelligence.<sup>2</sup>

### *1.5.2.1 Determinists and Artificial Intelligence*

The algorithmic objection need not produce a major problem to a determinist. To a determinist, everything is pre-determined. Blatchford writes:

To begin with, the average man will be against me. He knows that he chooses between two courses every hour, and often every minute, and he thinks his choice is free. But that is a delusion: His choice is not free. He can choose, and does choose. But he can only choose as his heredity and his environment cause him to choose. He never did choose and never will choose except as his heredity and environment—his temperament and his training—cause him to choose. And his heredity and his environment have fixed his choice before he makes it. ... There is a cause for every wish, a cause for every choice; and every cause of every wish and choice arise from heredity, or from environment.

For a man always acts from temperament, which is heredity, or from training, which is environment.<sup>3</sup>

In this view the behaviour of humans is just as deterministic as that of a computer, (this conclusion applying to both the hard and soft deterministic positions). People may be able to choose,

---

<sup>1</sup>Harel, David, *Algorithmics The Spirit of Computing*, Addison-Wesley Publishing Company, Wokingham, England, 1987, p. 336.

<sup>2</sup>Here the algorithm controlling the computer is presumed to be deterministic (see the next section). In this case, even processes which are highly complex (such as random number generation) are, to a philosopher, predictable.

<sup>3</sup>Blatchford, Robert, "The Delusion of Free Will", in Klemke, et. al., pps. 103, 104.

but the choice is just as fixed and determined as choices made by a computer running an algorithm.

For people holding this view, the objection that "computers just run algorithms" need not be relevant to the concept of computer intelligence.<sup>1</sup> It need not be an objection; and in fact could be argued as a confirming factor on the basis of similarity. This view would be relevant whether the person is a *materialist* (one who believes in the existence of only material entities) or a *dualist* (defined below).

### 1.5.2.2 Libertarians and Artificial Intelligence

For people holding libertarian views, whether the algorithmic objection can be decisive or not depends to at least some extent on another factor of the axiomatic views they hold.

Libertarians believe that "Matter cannot think".<sup>2</sup> They believe that there exists something extra, some *élan vital*, some life-force, some mind, soul or spirit; and hence postulate the existence of something other than the merely physical. *Idealists* believe "that only mind really exists and that matter is an illusion".<sup>3</sup> *Materialists* believe that only material states exist, that the *élan vital* in it's various guises is a non-existent state of affairs, and do not believe that a:

---

<sup>1</sup>E.G. Humphrey comments that 'These days the gurus of AI, such as Marvin Minsky ... do indeed simply take it for granted that there is nothing more to consciousness than sophisticated information processing - and talk blithely about robots that would be conscious merely by virtue of their ability to manipulate symbolic representations'; see: Humphrey, Nicholas, "The private world of consciousness", *New Scientist*, 8 January 1994, New Science Publications, London, 1994, pps. 23-25.

However Scriven approaches the topic from a somewhat different viewpoint when he comments 'I would now say it is now readily provable that the kind of free will required to make sense of the idea of responsibility and punishment is perfectly compatible with determinism and third-party predictability, and there is no evidence for any other kind. Hence, even if machines are predictable it would be possible for them to have free will.', Scriven, Michael, 'The compleat robot; A prolegomena to androidology', in Crosson, Frederick J., (Ed.), *Human and Artificial Intelligence*, Appleton-Century-Croft, New York, 1970, pps. 121-122.

<sup>2</sup>Taylor, Richard, "How to Bury the Mind-Body Problem", in Klemke, et. al., p. 178. Compare this with Weaver's comment 'it is no surprise that Shannon has just written a paper on the design of a computer which would be capable of playing a skilful game of chess. And it is of further pertinence to the present contention that this paper closes with the remark that either one must say that such a computer "thinks," or one must substantially modify the conventional implication of the verb "to think." ' see: Shannon and Weaver, pps. 25-26.

<sup>3</sup>Lycan, William G. (Ed.), *Mind and Cognition*, Blackwell, 1990, p. 3.



purely physical entity or state could have the property of being about or "directed upon" a non-existent state of affairs or object; that is not the sort of feature that ordinary, purely physical objects can have.<sup>1</sup>

Wooldridge suggests that even the presumably non-physical attributes of humanity should be examinable, hence we should:

accept the sense of consciousness itself as a natural phenomenon suited to being described by and dealt with by the body of laws and methods of the physical sciences. ... the property of consciousness is possessed only by very special organisations of matter (of types yet to be determined) when placed in a suitable electro-chemical state (that is still unknown).<sup>2</sup>

Crowson notes that:

T. H. Huxley and Ernst Haeckel alleged that life was just another property of organised matter, comparable to magnetism, electricity or heat—and as such should form the subject matter of an analytical science<sup>3</sup>

Sagan agrees with Huxley and Haeckel's materialist position, commenting:

My fundamental premise about the brain is that its workings—what we sometimes call 'mind'—are a consequence of its anatomy and physiology and nothing more.<sup>4</sup>

Sagan's position is, in this respect, in a classic tradition of materialist philosophy held by physicists; however Morowitz comments:

---

<sup>1</sup>Op. cit., p. 10.

<sup>2</sup>Wooldridge, Dean E., 'Computers and the Brain', in Crosson, Frederick J., (Ed.), *Human and Artificial Intelligence*, Appleton-Century-Croft, New York, 1970, pps. 76-78.

<sup>3</sup>Crowson, R. A., *Classification and Biology*, Heinemann Educational Books Ltd., London, 1970, p. 15.

<sup>4</sup>Sagan, Carl, *The Dragons of Eden*, quoted by Morowitz, Harold J., 'Rediscovering the Mind', in *THE MIND'S I*, p. 35.

Something peculiar has been going on in science for the past 100 years or so. Many researchers are unaware of it, and others won't admit it even to their own colleagues. But there is a strangeness in the air. What has happened is that biologists, who once postulated a privileged role for the human mind in nature's hierarchy, have been moving relentlessly towards the hard-core materialism that characterised nineteenth-century physics. At the same time, physicists, faced with compelling experimental evidence, have been moving away from strictly mechanical models of the universe to a view that sees the mind as playing an integral role in all physical events. It is as if the two disciplines were on fast-moving trains, going in opposite directions and not noticing what is happening across the tracks.<sup>1</sup>

Which of these fundamental views is held can influence a person's opinion of the possibility of the existence of artificial intelligence.

For materialists and idealists, there is no distinction between the fundamental constituent components making up machines and other objects or states. Considering only the algorithmic objection, this would seem to lessen the likelihood of objection to the existence of artificial intelligence.<sup>2</sup>

By contrast, the libertarian (sometimes loosely called a dualist) view asserts the following:

1. People are composed of two distinct and radically different entities—a body and a mind.
2. It is the body that is studied by physiologists, and it is the body that eventually rots.
3. Bodies are material entities, that is, entities which are essentially spatial or, less generally, entities which are describable by the fundamental properties of physics—distance, mass, etc..
4. It is the mind that thinks, feels, perceives, and meditates.

---

<sup>1</sup>Morowitz, Harold J., 'Rediscovering the Mind', in *THE MIND'S I*, p. 34.

<sup>2</sup>However there are considerations other than the algorithmic objection, especially in the categories of qualia and intentionality, see section 1.5.2 of this thesis.

5. The mind is a non-material entity, an entity which is not located in space.
6. The minds and bodies of human beings interact causally.<sup>1</sup>

The term *mind* is used in a number of ways by philosophers, Taylor comments that in this case "minds are entities which are radically different from bodies".<sup>2</sup> He then goes on to define varieties of dualism:

Technically, not every version of dualism asserts point 6. One version, parallelism, asserts that minds and bodies do not interact. Another version, epiphenomenalism, asserts that bodies affect minds causally but not vice versa. The most commonly discussed version, dualistic interactionism, is the view represented by the six points. According to this view, bodies affect minds and vice versa. Following common use, we refer to this view simply as dualism.<sup>3</sup>

Taylor's use of the term dualism is accepted in the following discussion, with the exception that the parallelist and epiphenomenalist positions will be discussed in section 1.5.2.2.4.<sup>4</sup> Christian, Aristotelian and pantheist (interactionist) dualist position with respect to the algorithmic objection will be

---

<sup>1</sup>Klemke, et. al., p. 160. For confirmation of point 6 see also Lycan, p. 3. Trusted refers to neurological experiments, and comments 'This is clear evidence that a mental event associated with a neural event precedes voluntary action. Other support that mental events are best regarded as taking hegemony and can be viewed as teleological (final) causes comes from the fact that subjects can be taught to control (produce or suppress) some of their brain activities..'; see Trusted, Jennifer, *Free Will and Responsibility*, Oxford University Press, 1984, p. 109.

<sup>2</sup>Klemke et. al., p. 161. Others concur. 'Sir John Eccles, who won the Nobel prize in 1963 for his studies of chemical communication between nerve cells ... with the philosopher Karl Popper ... has ... pursued a dualist course ... Popper and Eccles talk in terms of ... World 1 [which concerns] physical realities and World 2 equally real (but immaterial) mental states. ... [the laws of physics suggest that] Having no matter, the mind cannot influence the material brain. Eccles, however, now believes that the problem is solved by quantum mechanics. At the level of the very small, mere changes in the probability of an event ... can effect the functioning of the system as a whole. ... In the nervous system, the ideal site for such a mechanism is the synapse ... where tiny packets of chemical transmitter substances are released in a probabilistic fashion. Most current neurobiological thinking identifies the synapse as the place where changes take place that underlie learning and memory. The mind, Eccles ... concludes, exerts its effects by influencing the probability that packets of transmitter will be released' Ferry, p. 43.

<sup>3</sup>Klemke et. al., p. 161.

<sup>4</sup>Interactionist dualism is similar in emphasis to Cartesian dualism, see Lycan, p. 3.

discussed in sections 1.5.2.2.1, 1.5.2.2.2 and 1.5.2.2.3 respectively. The views are summarised in section 1.5.2.2.5

#### 1.5.2.2.1 Christian Libertarians

A Christian dualist could hold that the *élan vital* (soul) was restricted to mankind, since:

the Lord God formed man of the dust of the ground, and breathed into his nostrils the breath of life; and man became a living soul.<sup>1</sup>

If this immortal soul was a gift of God who restricted it to mankind, the ability to have free will would also be restricted to mankind.<sup>2</sup> Animals could thus not have a soul, mind or consciousness.<sup>3</sup> Machines which followed an algorithmic chain of events would be fundamentally different to mankind, being viewed as merely automatons, and hence not capable of intelligence as a theorem resulting from the basic axioms of Christian interactionist dualist belief, (intelligence being regarded as a consequent of the existence of a mind (soul, *élan vital*)).

---

<sup>1</sup>Genesis 2:9; *The Holy Bible*, Collins Clear-type Press, Glasgow, p. 16; (King James translation).

<sup>2</sup>Lovelock comments that René Descartes(1590-1650) also distinguished humans from all other living things in alone possessing a soul; (Lovelock, a proponent of the Gaia hypothesis, thinks Descartes wrong): see Lovelock, James, *Healing Gaia Practical Medicine for the Planet*, Harmony Books, New York, 1991, p. 31.

<sup>3</sup>This view, and/or its consequences, is still widely noted Begley & Ramo comment ' "Animal consciousness is still taboo." asserts ethologist Donald Griffin in his 1992 book "Animal Minds." ... Griffin says many science journals refuse to publish papers on the possibility of animal consciousness. ... "If you are a young and insecure scientist trying to get grants, a job or tenure, you would be ill advised to get into this," says the ethologist. "It is no coincidence that I did not get into it until I was not only tenured but almost retired." Part of the taboo stems from an insistence that only humans have minds. "When I began in the 1960s," recalls chimp biologist Jane Goodall, "you couldn't even ask about animal consciousness. [Even today,] there is strong pressure to make a distinction between us and the rest of the animal kingdom." Begley, Sharon & Ramo, Joshua Cooper, "Not just a pretty face", *The Bulletin with Newsweek*, 2 November 1993, ACP Publishing Pty. Ltd., Sydney, 1993, pps. 62-64. For another discussion of the views of Griffin (and others) see: Lewin, Roger, "I buzz therefore I think", *New Scientist*, New Science Publications, London, 15 January 1994, pps. 29-33. Also Vines attributes to Marion Stamp Dawkins the belief 'that birds and mammals at least experience forms of consciousness rather like ours' in her discussion of Dawkins' book *Through Our eyes Only? The search for animal consciousness* in: Vines, Gail, "The Emotional Chicken", *New Scientist*, 22 January, 1994, New Science Publications, London, 1994, pps. 28-31.

Dreyfus appears to hold views which seem to result in a similar importance being placed on the presence of humanity. Davidson comments:

The arguments Dreyfus uses have been called "situatedness"—intelligence has to be in a physical body, which, in turn, is situated in a particular culture in a particular time and place.<sup>1</sup>

If a test for machine intelligence (e.g. the Turing test)<sup>2</sup> was proposed and passed, this system of belief would not allow the acknowledgment of machine intelligence; it would simply take it as implicit that the test was incompetent, and propose a "more appropriate" test (e.g. the Chinese room)<sup>3</sup> which the proposer believed current machines could not pass. This process leads to what this author has called *the rainbow effect* where, regardless of the progress of artificial intelligence, the currently defined goal being sought always seems as far away as it ever was.<sup>4</sup>

Newquist recently provided an example of both the rainbow effect and the algorithmic objection when he wrote:

"If we could just get this machine to tell the difference visually between an apple and an orange, we would be giving it rudimentary intelligence" was a common thought in the industrial and academic sectors as recently as 10 years ago. ... Today, vision is not really considered AI, basically because ... algorithms have been developed that allow machines to do a

---

<sup>1</sup>Davidson, Clive, "Common sense & the computer", *New Scientist*, 2 April 1994, IPC Magazines Ltd., England, 1994, p. 33. This is similar to an extension of a terse view attributed to Minsky, that one neuron does not exhibit intelligence but 10<sup>11</sup> neurons do.

<sup>2</sup>Turing, A.M., 'Computing Machinery and Intelligence', *Mind*, 1950, Vol. LIX, pps. 433-460; see also discussion in Lycan, p. 4.

<sup>3</sup>Searle's ideas are presented, together with commentaries from other interested parties, forming an interesting discussion in: Searle, John R., 'Minds, Brains and Programs', *The Behavioural and Brain Sciences* (1980) 3, pps. 417-457. Also see Searle, John R., 'Minds, Brains and Programs', in Hofstadter, Douglas R. and Dennett, Daniel C., (Eds.), *THE MIND'S I*, Penguin Books, Harmondsworth, England, 1982, pps 353-382; see also Searle, John R., 'Is the Brain's Mind a Computer Program', *Scientific American*, Vol. 262 No. 1, January 1990, pps. 20-25. A cautiously opposing view is presented by Churchland, Paul M. and Churchland, Patricia Smith, 'Could a Machine Think', *Scientific American*, Vol. 262 No. 1, January 1990, pps. 26-31.

<sup>4</sup>For another explanation of this effect, see Humphrey, Nicholas, "The private world of consciousness", *New Scientist*, 8 January 1994, New Science Publications, London, 1994, p. 23.

good job of distinguishing one object from another ... now it's (almost) too easy.<sup>1</sup>

#### 1.5.2.2.2 Aristotelian Libertarians

Not all dualists accept the restriction of the *élan vital* to mankind. Bourbaki, considering thought to be an aspect of intelligence, notes that Searle would extend aspects of thought to systems with a biochemical component:

Searle contends that just like other biological activities such as digestion and photosynthesis, thought is intrinsically dependent on the biochemistry of its origin,<sup>2</sup>

Aristotle expresses a similar view most succinctly. Taylor comments that:

philosophers of the highest rank, such as Aristotle, have felt driven to say that all living things, vegetables included, must have souls (else how could they be *living* things?)<sup>3</sup>

Holders of this belief may feel that either all living beings, or some sub-group (e.g. dolphins, whales and/or chimpanzees)<sup>4</sup>

<sup>1</sup>Newquist III, Harvey P., "The Other Side of AI", in *AI EXPERT*, Volume 7, No. 3, March 1992, p. 50.

<sup>2</sup>Bourbaki, Nick, 'Turing, Searle, & Thought', *AI EXPERT*, Vol. 7, No. 5, July 1990, p. 55. Searle's work is further referenced in the early pages of the Neural Networks section of this thesis, see Appendix B.

<sup>3</sup>Taylor, Richard, "How to Bury the Mind-Body Problem", in Klemke, E.D. et. al., p. 177. An emphasis similar to Aristotle's has emerged recently in the "deep" ecology movement which 'places wildlife "as a fellow member of the moral community to which humankind belongs"'; Bagnall, Diana, 'New crimes of the times', *The Bulletin with Newsweek* Vol. 114, No. 5843, ACP Publishing Pty. Ltd., Sydney, 3 November 1992, p. 41.

<sup>4</sup>Taylor may be being kind to Aristotle when he attributes to Aristotle the view that 'all living things, vegetables included, must have souls...'. It is a matter of record that Aristotle's beliefs, as reported in his writings, only allowed the acceptance of what would now be classified as *a sub-group of living things* as having souls; Aristotle excluded women. Judith Brown comments 'In Aristotle's account of human generation, women are incomplete and imperfect males: "Just as it sometimes happens that deformed offspring are produced by deformed parents, and sometimes not, so the offspring produced by a female are sometimes female, sometimes not, but male. The reason is that the female is as it were a deformed male; and the menstrual discharge is semen, though . . . it lacks one constituent, and only one, the principle of Soul . . . . Thus the physical part, the body, comes from the female, and the Soul from the male, since the Soul is the essence of a particular body." *De generatione animalium*, II.3. 737a, 737b, trans. by A. L. Peck (Cambridge, Mass., 1943);' ; see Brown, Judith C., *Immodest Acts*, Oxford University Press, Oxford, 1986, p. 188.

Others have a broader view of the limits of life. Horgan comments 'the definition of "life" is becoming awfully flexible lately. Some computer scientists think a computer that simulates any living system, from a brain to a colony of algae, is a

exhibit some form of intelligence. "The word "animal" comes from a Latin root that means "soul."".<sup>1</sup> However they may not accept machine intelligence because they do not classify the computer as a *living* entity; and since "life had generally been thought of as a transcendental, God-given property, not amenable to scientific analysis"<sup>2</sup>, no amount of scientific inquiry or experimentation would be likely to find a way to make an inanimate object like a computer living. Again the existence of machine intelligence is likely to be rejected as the result of a theorem which is a consequence of the axioms of the observer's belief system.

#### 1.5.2.2.3 Pantheist Libertarians

Some philosophers take a less restricted view than Aristotle. Pantheists and animists (including some people accepting the Gaia hypothesis) see spirits in many or all things in the universe, (though generally preferring to allocate an *élan vital* exclusively to things they would see as being of "natural" origin).<sup>3</sup>

We should know the Great Spirit is within all things: the trees, the grasses, the rivers, the mountains, and the four-legged and winged peoples ...<sup>4</sup>

*To ancient thinkers, soul was the mysterious force that gave life and breath to the myriad of the earth's creatures. Some even*

---

kind of "artificial life." Julius Rebek, Jr., a chemist at the Massachusetts Institute of Technology, makes a more modest claim. He proposes that any assemblage of chemicals — not just ones consisting of proteins and nucleic acids — is arguably alive if it acts alive.' Horgan comments that 'Rebek refers to this area as "extrabiology", a term he [Rebek] has coined to describe the simulation of life in nonbiological systems. "Whether they involve synthetic molecules in vitro or computer constructs in silico ... these studies are intended to extend, then subsume that which is currently considered molecular biology." Watch out.', see: Horgan, John, 'Life in a Test Tube?', *Scientific American*, Vol. 266 No. 5, May 1992, p. 14.

<sup>1</sup>Kowalski, Gary, *The Souls of Animals*, Stillpoint Publishing, Walpole, U.S.A., 1991, p. 104.

<sup>2</sup>Crowson, p. 15.

<sup>3</sup>Interestingly, Cromer comments 'If there is one universal human characteristic ... it is a pervasive irrationality based on the egocentric confusion of self and other. ... Animism is the attribution of aspects of the self to objects and events. ... All children are animistic, and animism continues throughout life unless strongly controlled by contrary cultural guidance. It is universal among pre-literate peoples, who believe spirits inhabit all things - animate, inanimate and supernatural.' 'science ... is possible only after it is recognised that thought has no "real" power. For many, this final break with animism and magical thinking has been too high a price to pay for science.' quotations selected from Cromer, pps. 28-30.

<sup>4</sup>Elk, Black, quoted in Kowalski, p. v.

*spoke of a "world soul" or anima mundi that enlivened the whole of nature.*<sup>1</sup>

within Plato's anthropocentric universe ... the Cosmos itself was alive. ... The current interest in the defence of the environment and the attempt to establish legal rights for mountain, wilderness and recreation areas in, in a sense, a return to the Platonic view of the living Cosmos.<sup>2</sup>

Some modern writers, (again including some postulating the Gaia hypothesis) take up this ancient theme and describe the complete postulated system as an *organism*; for example:-

To a geophysicologist, a living organism is a bounded system open to a flux of matter and energy, which is able to keep its internal medium constant in composition, and its physical state intact in a changing environment; it is able to keep in homeostasis. ... This geophysicologist's definition of life includes Gaia. ... Gaia ... [is] ... able to regulate itself in a way like a living organism. ... I ... prefer the broad view that includes everything that metabolises and self-regulates as being alive, so that life is something shared in common by cats and trees, as well as beehives, forests, coral reefs, and Gaia. ... I respect the views of those with faith who find comfort in a Church, and who say their prayers, but acknowledge that they cannot, by logic alone, convince themselves, or others, of their reasons for believing in God. Similarly I respect those who take their comfort from the natural world and who may wish to say their prayers to Gaia.<sup>3</sup>

After attracting considerable criticism, Lovelock comments that he is willing to accept the term *living* as metaphoric in relation to Gaia.<sup>4</sup> However other pantheist and Platonic systems of belief may include inanimate components (usually of "natural" origin), and since these have been accepted as part of a living and/or spiritual system there is less reason to reject the concept of machine intelligence than in the other dualistic categories

---

<sup>1</sup>Kowalski, p. 104. The italics appeared in the original.

<sup>2</sup>Blakemore, Colin, *Mechanisms of the Mind*, Cambridge University Press, Cambridge, 1977, p. 13. See also the discussion in Gribbin, John, "Is the Universe alive?", *New Scientist*, New Science Publications, London, 15 January, 1994, pps. 38-40.

<sup>3</sup>Lovelock, p. 31.

<sup>4</sup>Lovelock, p. 6.



briefly mentioned above. Hence under this belief system, machine intelligence may be possible; but the emphasis on the putative importance of natural origins would seem to make this unlikely.

#### 1.5.2.2.4 *Parallelist and Epiphenomenalist Libertarians*

Both parallelist and epiphenomenalist libertarians accept the view that, although some sort of *élan vital* does exist, it does not directly influence material objects or things.<sup>1</sup> Thus the general position of people with these beliefs is similar to the positions discussed in sections 1.5.2.2.1 — 1.5.2.2.3, with the exception that the axiomatically accepted free will is demonstrated in purely materially determined behaviour, the *élan vital*, mind or soul having no influence in any of the decisions made. Whilst this point may appear to assist a bias in favour of the possible acceptance of the existence of artificial intelligence, the emphasis in sections 1.5.2.2.1 - 1.5.2.2.3 on the possessor of intelligence being *living* tends to make the effect of any such influence minimal, if it exists at all. The problem of having to classify potential possessors of artificial intelligence as living still remains. Few libertarians would classify a computer as living, except in jest. Hence, as a theorem from their axiomatically accepted beliefs, artificial intelligence can not exist.<sup>2</sup>

#### 1.5.2.2.5 *Summary; Libertarians and Artificial Intelligence*

In general, however, as Blatchford previously noted, the predominant view in modern (western) society is that mankind has, and can meaningfully exercise, free will. The implicit rejection of the concept of machine or artificial intelligence consequent to the acceptance of these (often unexamined)

---

<sup>1</sup>Regarding Parallelism, Kroy comments 'B. Spinoza, "The Ethics", in P. H. M. Elwes (ed., tr.), *The Chief Works of Benedict de Spinoza* (Dover, 1951), v. ii, p. 86 (Ethics pt. ii, prop. 7) says: "The order and connection of ideas is the same as the order and connection of things." This is the most concise version of the position I know.'; Kroy, p. 55.

<sup>2</sup>Perhaps whimsically one could suggest that, from this point of view, if a machine *could* hold an opinion about intelligence, it should be: 'Such knowledge is too wonderful for me; it is high, I cannot attain unto it'. (Psalm 139:6). However if a machine was wise enough to know it's own limitations, would this not be a form of intelligence?...

dualistic axioms may be behind some of the controversy as to whether artificial intelligence exists at all.

Certainly the *New Scientist* was not impressed with the idea. It stated that the '*New Scientist* does not believe in fairies, flying saucers or artificial intelligence'.<sup>1</sup> Similarly, Harel comments 'To many people the very idea of an intelligent machine does not sound right.'<sup>2</sup> Leary is more trenchant, '*Artificial intelligence* is an oxymoron'.<sup>3</sup>

The concept of a machine or artificial intelligence has been so strongly repugnant to some, that even the idea of using artificial intelligence as a modelling tool to help understand human intelligence has been strongly condemned:

'artificial intelligence as a way of understanding human behaviour ... is dehumanising and ideologically pernicious, undermining human agency and responsibility, and presenting a travesty of human potential'<sup>4</sup>.

Part of the reason for this repugnance may be distrust of a device that is not seen as having humanistic values. With no humanistic values as moderating influences, a fear of unacceptable decisions can exist:

In our culture an intelligent machine is immediately assumed to be a bad machine. As soon as intelligence is ascribed to the artificial, some people believe that the artifact will become evil and strip us of our humanistic values. Or, like the great gazelle and the water buffalo, we will be placed on reserves to be pampered by a ruling class of automata.<sup>5</sup>

---

<sup>1</sup>New Scientist, *New Scientist does not believe in fairies, flying saucers or Artificial Intelligence*, New Scientist, 8 November 1984, IPC Magazines Ltd., England, 1984.

<sup>2</sup>Harel, David, *Algorithmics The Spirit of Computing*, Addison-Wesley Publishing Company, Wokingham, England, 1987, p. 336.

<sup>3</sup>Quotation attributed to Timothy Leary in Ditlea, Steve, 'Artificial Intelligence', *Omni*, Volume 9, Number 7, Omni Publications International Ltd., New York, April 1987, p. 24. The italics were in the original article.

<sup>4</sup>Noted as a common view of A.I. in Kitzinger, Celia, 'Margaret Boden: Probing the mystery of the human mind', *The Psychologist*, Vol. 4, No. 1, January 1991, p. 14.

<sup>5</sup>Negroponte, Nicholas, quoted by Baecker, R.M., Buxton, W. A. S., "An Historical and Intellectual Perspective", in Preece, Jenny and Keller, Laurie (Eds.), *Human-Computer Interaction*, Prentice Hall, Hertfordshire, Great Britain, 1990, p. 19. Baecker and Buxton do not give the source of their quotation, but comment that

The quotations above express various varieties of unease in the concept of artificial intelligence.

It is argued that, however the rejection of artificial intelligence is couched, a rejection of the concept of artificial intelligence from a dualist is a consequence of the axiomatic beliefs inherent in that dualist position.<sup>1</sup>

### 1.5.3 Cognitive modelling and Artificial Intelligence

By contrast with the dualist position, the cognitive psychologist Anderson sees no problem in connecting algorithms and intelligence.<sup>2</sup> When writing about his cognitive theory of intelligence and development, he comments:

The central proposition of the theory is that intelligence is a property of thinking. In Chapter 5 I proposed that when someone is thinking, he or she is running an algorithm ...<sup>3</sup>

Anderson's brain-based cognitive model consists of a basic processing mechanism, and a series of specific processors.<sup>4</sup> The basic processor is postulated to provide the background ability, and the specific processors provide specific abilities which develop as the child ages.<sup>5</sup>

---

an elaboration of Negroponte's views can be found in Negroponte, Nicholas, *Soft Architecture Machines*, Cambridge, MA., the MIT Press, 1975.

<sup>1</sup>In fairness, it should be noted that the rejection of ideas that conflict with a person's basic beliefs and interests is not unique to the areas surrounding the ideas of artificial intelligence; witness "Rutherford's celebrated apophthegm 'all science is either physics or stamp-collecting'"; quoted in Crowson, R. A., *Classification and Biology*, Heinemann Educational Books Ltd., London, 1970, p. 10.

<sup>2</sup>As might be expected, Searle disagrees, arguing that 'the big mistake in cognitive science is not the overestimation of the computer metaphor (though that is indeed a mistake) but the neglect of consciousness'; Searle, John R. 'Consciousness, explanatory inversion, and cognitive science', *Behavioural and Brain Sciences*, 13, 1990, pps. 585-642; (an interesting commentary from other participants is included in this reference). A report of an attempt to investigate the idea of consciousness is given in: Crick, Francis and Koch, Christof, 'The Problem of Consciousness', *Scientific American*, Vol. 267 No. 3, September 1992, pps. 110-117. For a much more complete examination of consciousness see Edelman.

<sup>3</sup>Anderson, Mike, *Intelligence and Development A Cognitive Theory*, Blackwell Publishers, Oxford, 1992, p. 198. This view can be related to the "Man qua computer" metaphor, which Kroy states as 'Both Man and Computer have a "mind" (a system of programs) formally described and a "body" in which this mind is realized'; Kroy, p. 83.

<sup>4</sup>E.g. see Fig. 6.1, p. 107, Anderson, Mike.

<sup>5</sup>The development of specific abilities via the postulated specific processors could be used to explain Piaget's stages, see section 1.3.5 of this thesis.

If a computer employing currently-available levels of "artificial intelligence" was compared to a human adult using this model, the computer would probably be classified as an *idiot savant*.<sup>1</sup> Some of the specific processors would be reasonably well developed, others would be almost absent. If one accepts Gardner's six candidates for his multiple-intelligences theory, computers could possibly be rated highly in the logical-mathematical area, less well in the spatial area, not very well in the linguistic and musical areas, and hardly at all in the bodily-kinaesthetic and personal areas.<sup>2</sup> This imbalance would be regarded as markedly abnormal if it was observed in a human.

#### 1.5.4 Computer Science and Artificial Intelligence

Despite the philosophical problems already discussed, some computer scientists appear to accept the possibility of the existence of intelligence in a machine. An example occurs in the area of mobile autonomous robots; in 1993 there was a school provided by the NATO Advanced Study Institute entitled 'THE BIOLOGY AND TECHNOLOGY OF INTELLIGENT AUTONOMOUS AGENTS'.<sup>3</sup> The purpose of the Institute is listed as follows: 'The Advanced Study Institute brings together top-level researchers and practitioners from the emerging field of intelligent autonomous agents'.<sup>4</sup> The school includes a lecture by Luc Steels which 'focuses on how intelligence can be achieved in real world autonomous agents'.<sup>5</sup>

If artificial intelligence can be attained in machines, how can it be defined? Minsky gives a definition when he relates A.I. to human abilities, 'artificial intelligence is the science of making machines do things that would require intelligence if done by men'.<sup>6</sup>

---

<sup>1</sup>English and English (p. 250) define an *idiot savant* as "a feeble-minded person possessed of a high degree of some special ability, such as the ability to calculate".

<sup>2</sup>Anderson, Mike, *Intelligence and Development A Cognitive Theory*, Blackwell Publishers, Oxford, 1992, p. 67; or Evans, Peter and Deehan, Geoff, *The Descent of Mind*, Grafton Books, London, 1990, pps. 47-51.

<sup>3</sup>Gabora, Liane and Collins, Rob (Eds.), *Alife Digest*, Artificial Life Research Group, UCLA, Los Angeles, Volume #088, October 28th, 1992, p. 2.

<sup>4</sup>Ibid..

<sup>5</sup>Gabora, et. al., p. 5.

<sup>6</sup>Minsky, M. L., quoted in Boden, Margaret A., *Artificial Intelligence and Natural Man*, Basic Books, Inc., New York, 1977, p. 4; see also Aleksander, Igor, *Designing Intelligent Systems*, Billing & Sons Limited, Worcester, Great Britain, 1984, p. 18.

Boden makes 'no basic distinction between "artificial intelligence" and "computer simulation"'.<sup>1</sup> This approach fits with Simon's research group at Rand and Carnegie-Mellon University, who preferred the phrase *simulation of cognitive processes* to the term *artificial intelligence*.<sup>2</sup>

Simon defines the word *simulate* indirectly.<sup>3</sup> He comments that artificiality connotes perceptual similarity but essential difference, resemblance from without, rather than within. Simulation is possible because distinct physical systems can be organised to exhibit nearly identical behaviour; for example a damped spring and a damped circuit can both be described by the same second-order linear differential equation; hence one may be used to imitate or simulate the other. It is this sense of the word *simulate* which will be used when *artificial intelligence* is described as the simulation of human cognitive processes. Induction as discussed in this thesis will thus be restricted to being compared to human *cognitive* processes to at least partially side-step the philosophical problems discussed earlier relating to the concept of the *élan vital*, soul, mind or spirit.

The discussion in this thesis will also be restricted to the application of induction in the area of expert systems. Even the New Scientist regards expert systems as less controversial. In the week following the publication of their previously quoted rejection of artificial intelligence, an article was published in which Anderson quoted a former IBM scientist, Herbert Grosch, as saying:

AI - is stark naked from the ankles up. From the ankles down . . . [it] . . . is wearing a well worn and heavily-gilded [sic] pair of shoes called expert systems'<sup>4</sup>...

---

<sup>1</sup>Boden, Margaret A., *Artificial Intelligence and Natural Man*, p. 5.

<sup>2</sup>Simon, footnote p. 7.

<sup>3</sup>*ibid.*, p. 17.

<sup>4</sup>Anderson, Ian, "AI is stark naked from the ankles up", New Scientist, 15 November 1984, IPC Magazines Ltd., England, 1984.

## 1.6 Summary; induction, humans and expert systems

If the system proposed in the previous sections is correct, in the case of humans the faster responses are produced by inductive rather than deductive mechanisms. Deductive reasoning is used by those who have developed this facility when they have time for the use of this slower mechanism. The results of deduction can then be stored in memory, or associated with a situation by mental rehearsal, for later inductive access when a situation requires a further, prompt response.

In expert systems, the reverse applies. Relatively slow simulated inductive logic is used to examine raw data. Any relationships found are stored, (often in the form of implied decision trees) which can be later used by the relatively faster and more efficiently coded deductive logic.

When the strengths and weaknesses of both systems are considered, it can be seen that both methods sensibly use their fastest form of reasoning to provide best response in time-critical interactive situations.

It is this latter use of simulated logic, (use in an expert system), that is relevant to this thesis. Philosopher's views as to whether these forms of intelligent reasoning can be successfully implemented in a machine were then considered, and it was concluded that a person's view as to whether this is possible or not may depend on their basic philosophical axioms. A discussion of matters related to the computer implementation of these concepts follows.

# BACKGROUND TO COMPUTER SIMULATION OF INDUCTION

Artificial Intelligence systems which use induction are mostly classified under the general heading of learning systems. This chapter gives some background and history to various previous approaches in the area of artificial learning and associated systems, and notes some successes.

If one wishes to use the learning system approach, the first thing one must do is to reduce the voluminous data received by one's senses, some form of data compression technique being useful. Inductively formed keys are suggested as a suitable form of data compression; see section 2.1.

When attempting to obtain data there can be difficulty in getting an expert to express his or her expertise in the form of rules suitable for use in an expert system, and that in this case also, inductively formed keys have proven useful, but with some disadvantages; see section 2.2.

Section 2.3 suggests that the aim of any new implementation of an inductive learning system should be to reduce some of these attendant disadvantages.

## 2.1 Deriving Rules to Systematise Data.

If one wishes to use the learning system approach, some systemisation of the environment is important. The first thing one must do is to reduce the impressions received by one's senses to some sort of numeric or categoric form; see section 2.1.1.

It is often difficult to spot trends, and form overall impressions from the resulting masses of numbers and categories; in this case data compression can be useful. One method of data compression often used in the identification of botanical data is the dendritic tree (key); see section 2.1.2.

There have been various approaches used by authorities in constructing keys, many employing some variation of Shannon's

re-discovered entropy function; see section 2.1.3. This entropy function allows the use of real number characteristic measurement. Not all methodologies use real numbers directly; in some methods the real numbers are classified into categories. Some of these category-using methodologies are noted in section 2.1.4. While categorised real numbers could be employed in the production of keys, the more powerful early methods generally produced better results when used with discrete data; see section 2.1.5.

### 2.1.1 Discrete and Continuous Data

There is evidence the information a person receives from the environment is systematised and summarised by the neural circuits receiving the information before it is examined by the brain. Certainly this general approach is of use when attempting to handle large amounts of data which may have been gathered about objects which have been under study.

In humans, this data would be the result of an integration of impressions from the various senses.

When studying objects, the data would consist of categorised and measured characteristics associated with the objects being studied. The list of data may be variable in size and composition, but with computers, a data list is generally of fixed maximum length and composition, resulting from the formalised impression of a human or machine. It will be noted that data characteristics are generally occurrences of either of two forms:-

a) Numeric

e.g. length = 2.7, 4.0, 25, 9.713

b) A member or members of a set,

e.g. tint = red, blue, white, black;

These two types of data have historically been treated differently.

Consider the real number data. The amount of this data may be large, and some means of reducing it or quantifying it into a



more manageable form is often required. Statistical methods are one method of achieving this compression.

A second method of reducing the data to a manageable size is to categorise the statistical data. Categorical data can be treated by methods employing concepts such as information theory.<sup>1</sup>

### 2.1.2 Data Compression applied to real-numbered characteristics.

Historically, in an attempt to make trends more easily apparent, data has been represented as graphs and various types of diagrams. However, a breakthrough came when Galton examined the huge quantity of data he collected as part of his study of hereditary genius.<sup>2</sup> To assist in parameterising the data, he, with the aid of his students, employed and developed what has been called the handmaiden of the observational sciences - statistics. Basing his approach on previous work by Carl Friedrich Gauss, he parameterised the distributions in terms of an average and a measure of spread. Using this approach, he was able to show that this type of parameterisation could be used to accurately describe the chest measurements of (e.g.) 5,738 Scottish soldiers, and the heights of 100,000 French conscripts.<sup>3</sup> In this way the distributions of measurements could be replaced by a few parameters, a process Cohen and Feigenbaum call data compression.<sup>4</sup> Data compression made the distributions mathematically manipulable, and Gower notes that it was an associate of Galton, Karl Pearson, who published what seems to be the first paper on what would now be called statistical induction.<sup>5</sup>

---

<sup>1</sup> See also section 2.1.5 of this thesis.

<sup>2</sup> Galton, Francis, *Hereditary Genius*, Collins, London, U.K. 1962; (a reprint of the text and diagrams of Galton's second edition of *Hereditary Genius*, published by Macmillan & Co. Ltd., 1869).

<sup>3</sup> Galton, pps. 70-71.

<sup>4</sup> Cohen, Paul R., & Feigenbaum, Edward A., *The Handbook of Artificial Intelligence*, Vol. 3, HeurisTech Press, Stanford, California, 1982, p.383.

<sup>5</sup> Pearson, Karl, *Contributions to the mathematical theory of evolution. I. Dissection of frequency curves*, Phil. Trans. R. Soc., A 185, 1894, pps. 71 - 110, referenced in Gower, J. C., 'Relating Classification to Identification', in Pankhurst, R. J. (Ed.), *Biological Identification with Computers*, Systematics Association Special Volume No. 7, Academic Press, London, 1975, pps. 251 - 263.

If real-valued data is to be represented in key format, the data must be categorised. Categorised data is considered in the section 2.1.4.

### 2.1.3 Key Building - A Background History

Keys used for the identification of botanical species have long been constructed by hand. The development of appropriate mathematical methods meant that an algorithmic approach was also possible. Several methodologies have been used for key construction.<sup>1</sup> The most prominent methodology addressed in artificial intelligence publications in recent years has used an information function. Kullback comments:

Information in a technically defined sense was first introduced in statistics by R. A. Fisher in 1925 in his work on the theory of estimation. ... Shannon and Weiner, independently, published in 1948 works describing logarithmic measures of information for use in communication theory. ... it can be as is applied in a wide variety of fields.<sup>2</sup>

The similarity between information and entropy was noted by Szilard in a 1929 paper<sup>3</sup>, which was a forerunner of Shannon's paper.<sup>4</sup>

At the time, applications of information functions were perceived to be legion. Macnaughton-Smith shows the potential breadth of applications in the area of classification alone when he commented (regarding classification techniques) that:

workers in many fields have developed numerical techniques, which are scattered throughout the literature of Bacteriology, Botany, Ecology, Information theory, Microbiology, Philosophy, Zoology and other subjects.<sup>5</sup>

---

<sup>1</sup>For a summary of some early work in the area of key building during the 1960s see: Pankhurst, (1970a), pps. 147-148.

<sup>2</sup>Kullback, Solomon, *Information Theory and Statistics*, John Wiley & Sons, New York, 1959, p. vii.

<sup>3</sup>Szilard, Von L., 'Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen', in *Zeitschrift für Physik*, Volume 53, Verlag von Julius Springer, Berlin, 1929, pps. 840-856.

<sup>4</sup>Shannon, Claude E., 'A mathematical theory of communication', Bell Systems Technical Journal, Volume 27, 1948, pps. 379-423, 623-656.

<sup>5</sup>Macnaughton-Smith, P., *Some Statistical and Other Numerical Techniques for Classifying Individuals*, Her Majesty's Stationery Office, London, 1965, p. 1.

R. Quastler was referenced as editing of a volume of papers on the application of information theory in the biological area in 1953, including one by the editor on 'The measure of specificity'.<sup>1</sup>

H. Quastler comments:

The summer of 1954 saw at least three gatherings of people interested in the application of information theory to psychology. The Fourteenth International conference of Psychology was held in Montreal from June 7 to 12; on its agenda was a symposium on information theory arranged by J. C. R. Licklider. In the following week, a three-day conference on information theory was held at the Massachusetts Institute of Technology; this was also arranged by Dr. Licklider. On June 5-9, a five-day conference was held at Allerton Park; this conference was arranged by the Bio-Systems Group of the Control systems Laboratory at the University of Illinois. ... To anybody who needs an introduction to the field, we recommend George A. Miller's article "What is information measurement?" (The American Psychologist 8: 3, 1954).<sup>2</sup>

Authors at this conference reference papers which appear to be uses or discussions of the application of information measures in the life sciences going back to 1951. In one of the papers presented at that conference, Cronbach gives a series of conditions which he states should apply if Shannon's formulation is to be used unchanged. He also cautions that:

Shannon's "continuous case" makes no use of the fact that numbers are ordered. ... Psychologists who use Shannon often

---

<sup>1</sup>Quastler, R., 'The measure of specificity', in Quastler, R (Ed.), *Information Theory in Biology*, University of Illinois Press, Urbana, 1953 (not seen), referenced in Margalef, D. R., *General Systems*, Volume 3, p. 36, 1958. However the initial attributed to Margalef's usage by his translator is probably in error, as William J. McGill references a paper of the same name in a volume of identical title as being by H. Quastler, see: Quastler, H., 'The measure of specificity' (not seen), in Quastler, H (Ed.), *Information Theory in Biology*, University of Illinois Press, Urbana, 1953 (not seen), referenced by McGill, William J., 'Isomorphism in Statistical Analysis', in Quastler, Henry (Ed.), *Information Theory in Psychology*, The Free Press, Glencoe, Illinois, p. 62.

<sup>2</sup>Quastler, Henry, (Ed.), *Information Theory in Psychology*, The Free Press, Glencoe, Illinois, 1955, p. v-vi.

treat ordinal or interval data, and are thereby likely not to take full advantage of their data.<sup>1</sup>

In Australia at this time, Goodall was publishing a paper on the use of factor analysis in the examination of Australian botanical species.<sup>2</sup>

Garner and McGill continue the investigation of the application of information theory to the life sciences in their 1956 paper, noting that 'Psychologists have been attracted by the non-metric character of this measure and the obvious application to situations where variances cannot be computed. ... We shall show that uncertainty has many of the properties of variance and can be partitioned into components as variance can'.<sup>3</sup> (McGill and Quastler had earlier attempted to standardise the nomenclature in this field, and defined the measure 'uncertainty'.<sup>4</sup> However some later writers talk of 'entropy', using the 'remarkable likeness between information and entropy. This similarity was noticed long ago by L. Szilard,<sup>5</sup> in an old paper of 1929, which was the forerunner of the present theory. ... We prove that information must be considered as a negative term in the entropy of a system; in short, information is negentropy'.<sup>5</sup>) This reference also discusses information and computers.<sup>6</sup>

Margalef in his 1957 presentation discussed the use of Shannon's information measure in a study of the diversity of life forms, and the ability to distinguish between different species;

---

<sup>1</sup>Cronbach, Lee J., 'On the non-rational application of information measures in psychology', in Quastler, Henry, (Ed.), pps. 23-24.

<sup>2</sup>Goodall comments: 'Factor analysis does not result in a classification of vegetation in the ordinary sense, but in arrangement of the vegetational data in a multi-dimensional series. For such an arrangement, there appears to be no word in English which one can use as an antonym to "classification"; I would like to propose "ordination"; see: Goodall, D. W., 'Objective Methods for the Classification of Vegetation', Australian Journal of Botany, Volume 2, Number 1, Commonwealth Scientific and Industrial Research Organisation, East Melbourne, February 1954, p. 323.

<sup>3</sup>Garner, W. R. and McGill, William J., 'The relation between information and variance analysis', *Psychometrika*, Volume 21, No. 3, September 1956, p. 220.

<sup>4</sup>McGill, William and Quastler, Henry, 'Standardised Nomenclature: An Attempt', in Quastler, Henry (Ed.), 1955, pps. 83-92.

<sup>5</sup>Szilard, Von L., 'Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen', in *Zeitschrift für Physik*, Volume 53, Verlag von Julius Springer, Berlin, 1929, pps. 840-856. This reference has been added to the original quotation.

<sup>5</sup>Brillouin, Leon, *Science and Information Theory*, Academic Press, New York, 1956, p. xi-xii.

<sup>6</sup>See Brillouin, Chapter 19.

'Information theory describes the evolution of structured systems, divisible into elements qualitatively different'.<sup>1</sup>

In a paper published in 1960, Williams and Lambert followed up a 1959 paper (in which they hand calculated hierarchical diagrams by hand) with a paper which described the methodology and results achieved by:

a fully automatic programme ... for a Ferranti 'Pegasus' digital computer, capable of dealing with up to 76 species and either 1680 or 3200 quadrats, depending on the type of drum available.<sup>2</sup>

A quadrat was (vary roughly) a unit of data collected. They note that the Pegasus had a 38 bit word, and they used a bit to represent the presence or absence of a species, for 76 species two words could be used. A run of 76 species with 100 quadrats took 'something over an hour'.<sup>3</sup> This system used an association-index (rather than an information-based) methodology to form the hierarchical diagrams.

Macnaughton-Smith's 1965 Home Office publication discusses the use of several methodologies, including Shannon's information measure, for classification purposes.<sup>4</sup>

An early computer-based uses of Shannon's measure for classification was reported by Seshu in 1965, (with a possibility that Seshu & Freeman used this measure in a program reported in 1962).<sup>5</sup>

---

<sup>1</sup>Margalef, D. Ramon, 'Information Theory in Ecology', in *General Systems*, Volume 3, p. 36, 1958. p. 68. This paper was originally in Spanish and was presented by the author to the Royal Academy of Sciences and Arts of Barcelona on the occasion of his acceptance of election to the Academy on April 4, 1957. The English-language version was translated by Wendell Hall from *Memorias de la Real Academia de Ciencias y Artes de Barcelona*, 23: 373-449, November, 1957.

<sup>2</sup>Williams, W. T. and Lambert, J. M., 'Multivariate Methods in Plant Ecology', *The Journal of Ecology*, Volume 48, Blackwell Scientific Publications, Oxford, 1960, p. 689.

<sup>3</sup>*Idem.*, p. 696

<sup>4</sup>Macnaughton-Smith, P., *Some Statistical and Other Numerical Techniques for Classifying Individuals*, Her Majesty's Stationery Office, London, 1965.

<sup>5</sup>The program was written in machine language on a CDC-1604. It was used for diagnosis of machine failures. Sechu, Sundaram, 'On an Improved Diagnosis Program, *I.E.E.E. Transactions on Electronic Computers*; February 1965, Vol. EC-14, pps. 76-79. Sechu comments in this article that he had participated in writing a similar program (for an IBM 7090) earlier, but that that program had been less flexible and was proprietary. His 1962 article does not reveal the basis for the diagnosis, but if it was the same, this may have been the earliest recorded use of information gain (entropy) in a computer program for diagnosis. See

With regard to computer-based methodologies (not necessarily entropy-based), Gower notes that:

Working independently, Sneath(1957), Sokal and Michener (1958) and Williams and Lambert (1959), with interests in respectively in bacteriology, entomology and ecology produced computer programs for classificatory purposes and mostly for hierarchical classification. ... it stimulated widespread interest and encouraged many others to develop similar computer programs.<sup>1</sup>

The number of papers published in this area grew large, and only a few will be referenced in the following discussion. Pielou in 1966 reported results obtained with Brillouin and Shannon's measures of information, mainly looking at diversity of species in different types of biological collections. They also mention seven previous papers concerned with information content and diversity, (only one of which is cited in this discussion).<sup>2</sup>

In 1966 Williams, Lambert and Lance published a paper which included (amongst others) consideration of information measures in the field of plant ecology. It is of interest that, at the time the paper was published, both the first and last named authors were working at the C.S.I.R.O. in Canberra.<sup>3</sup>

Hunt et. al. report experiments in a 1966 publication that they used measures involving similarity of attributes as cost criteria for constructing binary trees in several versions of their Concept Learning System CLS.<sup>4</sup> Hunt comments that the 'concept

---

Seshu, S., and Freeman, D. N., 'The diagnosis of asynchronous sequential switching systems', *IRE Trans. on Electronic Computers*, Vol. EC-11, August 1962. pps. 459-465.

<sup>1</sup>Gower, J. C., 'Relating Classification to Identification', in Pankhurst, R. J. (Ed.), *Biological Identification with Computers*, Systematics Association Special Volume No. 7, Academic Press, London, 1975, pps. 253.

<sup>2</sup>Pielou, E. C., *The Measurement of Diversity in Different Types of Biological Collections*, *Journal of Theoretical Biology*, Volume 13, 1966, pps. 131-144.

<sup>3</sup>Williams, W. T., Lambert, J. M. and Lance, G. N., 'Multivariate Methods in Plant Ecology', *Journal of Ecology*, Volume 54, 1966, pps. 427-445.

<sup>4</sup>Hunt, E. B., Marin, J. and Stone, P. T., *Experiments in Induction*, Academic Press, New York, 1966. Note that in Cohen & Feigenbaum, p. 406, Cohen & Feigenbaum give CLS the position of being a precursor to ID3 when on p. 408 they describe step 2 of the ID3 algorithm as 'Use the CLS algorithm to form a rule to explain the current window'. Chapter 2 of Hunt, Marin & Stone gives a description of the methodology used by Hunt et al. Also see: Muggleton, Stephen, *Inductive Acquisition of Expert Knowledge*, Addison-Wesley, England, 1990, p. 173.

learning computation [used] is one of those proposed by Bruner et al. (1956) as algorithms for solving conjunctive learning problems'.<sup>1</sup>

Sneath and Sokal also discuss the use of entropy in classification<sup>2</sup>. The authority they quote is Orlóci's paper on information theory and hierarchical and non-hierarchical classification which was presented to a conference at the University of St. Andrews in Scotland in 1968.<sup>3</sup>

Orlóci in 1968 discusses the application of Brillouin's measure, Stirling's approximation and Shannon's measure of information to partition and classification in Phytosociological applications.<sup>4</sup>

Gower and Barnett in 1971 report using (amongst other methodologies) an extended version of Shannon's information criteria (entropy) to construct a botanic key which was subsequently tested by using it to identify specimens consisting of 68 species of fruit yeasts.<sup>5</sup> They refer to three other papers in dated 1968 to 1970 discussing the use of binary keys, but comment that they (Gower and Barnett) 'also consider unknown responses, denoted by "?"'.<sup>6</sup>

Gower and Payne extended the study in a paper published in 1975, in which five methodologies (including entropy) were compared. They commented that Pankhurst in 1970 suggested a form of the entropy function that would bias against tests with more than two responses, and rate the entropy function as the worst of the five methodologies considered for type 1 and type 2 errors, but the only one capable of unconditional extension to multi-response tests. Whereas all the other methodologies were

---

<sup>1</sup>Hunt et al., (1965), p. 21. The Bruner reference (not seen) is included in the bibliography for the sake of completeness.

<sup>2</sup>Sneath, Peter H. A. and Sokal, Robert R., *Numerical Taxonomy, The Principles and Practice of Numerical Classification*, W. H. Freeman and Company, San Francisco, 1973, p. 141-145.

<sup>3</sup>Orlóci, László, 'Information theory models for hierarchical and non-hierarchical classifications', in Cole, A. J. (Ed.), *Numerical Taxonomy*, Proceedings of the Colloquium in Numerical Taxonomy Held in the University of St. Andrews, September 1968, pps. 148-164, Academic Press, London.

<sup>4</sup>Orlóci, László, 'Information Analysis in Phytosociology: Partition, Classification and Prediction', *Journal of Theoretical Biology*, Volume 20, 1968, pps. 271-284.

<sup>5</sup>Gower, J. C. and Barnett, J. A., 'Selecting Tests in Diagnostic Keys with Unknown Responses', *Nature*, Vol. 232, August 13<sup>th</sup> 1971, pps. 491-493.

<sup>6</sup>*Ibid.*, p. 492.

rated as computationally efficient, entropy alone was rated as inefficient.<sup>1</sup>

Dunn and Everitt in 1982 comment:-

Except for dynamic programming algorithms, which effectively enumerate all possible keys (Garey, 1972),<sup>‡</sup> no exact algorithm is known for finding optimum keys. Dynamic programming algorithms, however, are impracticable for most real data, which may be concerned with several hundred taxa and, in some cases, of the order of a hundred characters<sup>#</sup> or more. Several authors, for example, Pankhurst (1970)<sup>§</sup>, Morse (1971), and Payne (1975) present algorithms giving approximate solutions. They all operate by selecting first the test that 'best' divides the taxa into two sets. Various criteria, some of which are described below, have been used to define what is meant by the best test. After the first division, the chosen criterion is used to select next test to be used with each subset of taxa, and so on. Garey & Graham (1974) give examples showing that selecting tests in this way, without examining their later consequences, can lead to inefficient keys, but most authors claim that their algorithms work well in practice and certainly give keys as good as, if not better than, those prepared by hand using intuition and experience.

For tests which have equal costs and taxa and for which there are not variable responses, the most common criterion used to

---

<sup>1</sup>Gower, J.C., and Payne, R. W., 'A comparison of different criteria for selecting binary tests in diagnostic keys', in *Biometrika*, Vol. 62 No. 3, 1975, pps. 665-672.

<sup>‡</sup>Note that, although no exact algorithm for finding an optimum key is known, the problem is known to be NP complete.

For the sake of completeness, these references have been reproduced in the reference list included in this thesis. (This footnote is not in the source quoted.)

<sup>#</sup> N.B. character = characteristic. In botanic literature, a 'character' is some characteristic of the specimen which can be measured or categorised, having (in this case) potential for subsequent use in the identification of the specimen. I have not used the term 'character' in this thesis, preferring 'characteristic' to avoid any potential confusion with the computer science use of the term 'character' (meaning a letter, number or punctuation mark etc. used in printing). (This footnote is not in the source quoted.)

<sup>§</sup> The reference given in Dunn & Everitt is: Pankhurst, R. J., *A computer program for generating diagnostic keys*, New Phytologist, Vol. 62, pps. 35 - 43, 1970; this reference could not be traced and Pankhurst commented 'The reference ... is completely fictitious! Presumably the Computer Journal paper is intended', (private communication). The Computer Journal reference is: Pankhurst, R.J., 'A computer program for generating diagnostic keys', *Computer Journal* Vol. 13 No. 2, May 1970a, pps. 145-151. (This footnote is not in the source quoted.)



choose the best test is based on the entropy function of Shannon (1948) and is given by

$$H_i = \sum_{k=1}^{m_i} p_{ik} \log p_{ik}, \quad (7.1)$$

where  $p_{ik}$  is the proportion of taxa with fixed response  $k$  to test  $i$  and  $m_i$  is the number of levels of test  $i$ . At each stage the test with minimum value of  $H_i$  is chosen.

For taxa having variable responses, Shwayder (1971, 1974) suggested that  $H_i$  be modified to

$$H'_i = H_i - (1 - r_i) \log(1 - r_i), \quad (7.2)$$

where  $r_i$  is the proportion of taxa with variable responses to test  $i$

A further function, suggested by Brown (1977), is

$$M_i = - \sum_{k=1}^{m_i} p_{ik} (1 - p_{ik} - r_i) \quad (7.3)$$

Gower & Payne (1975) investigated the properties of several such criteria and in the multiresponse case found the criterion

$$S_i = \sum_{k=1}^{m_i} (p_{ik} + r_i) \log(p_{ik} + r_i) \quad (7.4)$$

to be most suitable.<sup>1</sup>

When Dunn and Everitt noted in their 1982 publication that Shannon's entropy function equation 7.1 was (given the certain considerations) amongst the criteria most commonly used for key construction, they used a slightly different form of the function to that used later by Quinlan,<sup>2</sup> in that the form used omitted the minus sign used by Shannon.<sup>3</sup> The different form may seem a point for concern, but Weaver comments:-

<sup>1</sup>Dunn, G., and Everitt, B. S., *An introduction to mathematical taxonomy*, Cambridge University Press, Cambridge, 1982, pps. 110-111. Note that for the sake of completeness, the references noted in this quotation have been reproduced in the reference list included in this thesis.

<sup>2</sup>Quinlan, J. Ross, *Induction of Decision Trees*, Technical Report 85.6, School of Computing Sciences, New South Wales Institute of Technology, Sydney, 1985. Gower and Payne also note prior discussion of this point by Pankhurst in a 1970 publication.

<sup>3</sup>For the derivation of Shannon's entropy function, see Appendix 2 of Shannon and Weaver, pps. 116-117.

Do not worry about the minus sign. Any probability is a number less than or equal to one, and the logarithms of numbers less than one are themselves negative. Thus the minus sign is necessary in order that  $H$  be in fact positive.<sup>1</sup>

Dunn and Everitt's comment that Gower and Payne's 1975 survey of several functions of this type found that there may be problems with Shannon's entropy function in the case where there were multiple responses was confirmed by Quinlan in his 1985 technical report.<sup>2</sup>

Gower noted the computational problems inherent in the statistical approach were formidable, and 'the multivariate case can scarcely be tackled without modern computers'.<sup>3</sup> He reviews categoric and probabilistic approaches, noting 'when populations do overlap, a full probabilistic treatment is necessary'.<sup>4</sup>

Edgington in 1969 compared a variety of statistical approaches, including randomisation tests which make no assumption about the shape of the distribution, and include the idea of using a window or sub-set of the training data to increase computational efficiency.<sup>5</sup>

Quinlan in 1979 combined the entropy measure with a windowing concept to make calculation of the information gain (entropy) associated with decisions more computationally efficient, and applied this to the development of inductively generated keys in 1985. He called his algorithmic approach ID3.<sup>6</sup> Wirth et. al. report on experiments on the costs and benefits of the windowing that ID3 performs.<sup>7</sup>

---

<sup>1</sup> See Shannon and Weaver, p. 15.

<sup>2</sup> Gower, J.C., and Payne, R.W., *A comparison of different criteria for selecting binary tests in diagnostic keys*, in *Biometrika*, Vol. 62 No. 3, 1971, p.671.

<sup>3</sup> Gower, p.252.

<sup>4</sup> Gower, p.261.

<sup>5</sup> Edgington, Eugene S., *The Distribution-free Approach*, McGraw-Hill Book Company, New York, 1969, p. 152.

<sup>6</sup> Quinlan, J. Ross, *Induction of Decision Trees*, Technical Report 85.6, School of Computing Sciences, New South Wales Institute of Technology, Sydney, 1985.

<sup>7</sup> Wirth, Jarryl and Catlett, Jason, 'Experiments on the Costs and Benefits of Windowing in ID3', in Laird, John (Ed.), *Proceedings of the Fifth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Mateo, U.S.A., 1988, pps. 87-99.

Cheng et. al. claim there are several problems with the ID3 approach, namely:-

ID3 is essentially employing a heuristic, hill-climbing, non-back-tracking search through the space of possible decision trees. Thus, weaknesses in the ID3 algorithm may cause it to "miss" better decision trees for the same data. ... Perhaps the most pronounced is the **irrelevant values problem**. When ID3 chooses an attribute for branching out from a node, it creates a branch for each value of that attribute that appears in the examples. Some of the values of that attribute may be relevant to the classification, yet the rest may not be. ... Another related problem is the **missing branches problem**. The missing branches problem is essentially a reduction in the inductive capacity of the tree. It is due to the fact that some of the reduced sub-sets at the non-leaf nodes do not necessarily contain examples of every possible value of the branching attribute.<sup>1</sup>

Several improvements have been made to the general CLS and ID3 approach, e.g. see Cheng et. al., Cestnik et. al., Quinlan, Catlett and Collier.<sup>2</sup>

Also Brieman *et al.* have implemented a Classification And Regression Trees (CART) approach.<sup>3</sup> They comment:

---

<sup>1</sup>Cheng, Jie, Fayyad, Usama M., Irani, Keki B. and Qian, Zhaogang, 'Improved Decision Trees: a Generalised Version of ID3', in Laird, John (Ed.), *Proceedings of the Fifth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Mateo, U.S.A., 1988, pps. 100-106. The bold type is as preferred by the authors of the paper.

<sup>2</sup>Cheng et. al, see preceding footnote; Cestnik, Bojan, Kononenko, Igor and Bratko, Ivan, 'Assistant 86: A Knowledge-Elicitation Tool for Sophisticated Users', in Bratko, I. and Lavrac, N., *Progress in Machine Learning*, Sigma Press, England, 1987, pps. 31-45. Also Quinlan, J. Ross, 'Decision Trees as Probabilistic Classifiers', in Langley, Pat (Ed.), *Proceedings of the Fourth International Workshop on Machine Learning*, June 1987, Morgan Kaufmann Publishers, Inc., San Mateo, U.S.A., 1987, pps. 31-37; and Quinlan, J. Ross, 'Learning with Continuous Classes', in Adams, Anthony and Sterling, Leon (Eds.), *Proceedings of the 5<sup>th</sup> Australian Joint Conference on Artificial Intelligence*, World Scientific Publishing Co., Singapore, November 1992, pps. 343-348. Also Catlett, J., 'Peepholing: choosing attributes efficiently for megainduction', in Sleeman, Derek and Edwards, Peter, *Machine Learning Proceedings of the Ninth International Workshop*, Morgan Kaufmann Incorporated, 1992, pps. 49-54. Also Collier, P. A., *Manual for TL*, unpublished manuscript.

<sup>3</sup>Breiman, Leo, Friedman, Jerome H., Olshen, Richard A., and Stone, Charles J., *Classification and Regression Trees*, Wadsworth International Group, Belmont, California, 1984.

Many different criteria can be defined for selecting the best split at each node. As noted, in the ship classification project, the split selected was the split that most reduced the node impurity defined by

$$i(t) = -\sum_j p(j|t) \log[p(j|t)].$$

... Two splitting rules are singled out for use. One rule uses the *Gini Index of diversity* as a measure of node impurity; i.e.,

$$i(t) = -\sum_{i \neq j} p(i|t)p(j|t).$$

The other is the *twoing rule*: At a node  $t$ , with  $s$  splitting  $t$  into  $t_L$  and  $t_R$ , choose the split  $s$  that maximises

$$\frac{P_L P_R}{4} \left[ \sum_j |p(j|t_L) - p(j|t_R)| \right]^2 \quad 8$$

They note that with the CART methodology, the final tree selected is 'surprisingly insensitive to the choice of splitting rule'.<sup>1</sup> They also examine the 'missing value' problem, and implement *surrogate splits* to handle this.<sup>2</sup> They attempt to minimise the computational load by using subsampling if the class is bigger than a fixed maximum.<sup>3</sup> They claim that when compared to other methodologies with regard to accuracy, the CART results have 'generally been either best or close to best'.<sup>4</sup>

However fundamental problems with these general approaches still remain when these methodologies are applied to some complex problems both inside and outside the field of key generation. Some of these problems are discussed further in sections 2.2.2. and 2.2.3.

In addition, the computational load was still considered excessive, and other ways of handling this data, such as categorisation, were investigated.

---

<sup>8</sup>*Idem*; p. 38. The ship classification project is discussed in pps. 19 & 20 of this reference.

<sup>1</sup>*Ibid.*

<sup>2</sup>Breiman *et al.*, pps. 40, 248-251. Note that surrogate splits apply to only a single missing attribute value at a single test in the decision tree.

<sup>3</sup>*Idem* pps. 42, 163-167.

<sup>4</sup>*Idem* p. 171.

### 2.1.4 Categorisation of Numeric Data

Michalski & Chilausky reported an attempt to obtain rules (for soy-bean diseases), and some of their data was numeric.<sup>1</sup> They treated the numeric data by categorising it, (e.g. [precipitation<n]) and obtained results which 'were viewed generally quite favourably by experts - with a few exceptions'.<sup>2</sup>

By contrast, Williams reports experiments involving categorisation of real or integer data in an attempt to improve ID3's classification efficiency.<sup>3</sup> Generally the results reported are discouraging. This confirms unpublished results obtained by P. A. Collier in which far less satisfactory decision trees were obtained using categorised number data than were obtained with *1<sup>st</sup> Class* using uncategorised number data.<sup>4</sup> These results are not surprising as categorisation reduces the amount of information available to the inductive process.

### 2.1.5 Classification of discrete valued characteristics.

Better progress was made with algorithms for inductive classification of variables which had discrete options.

Systems which inductively classify the data fall under the general heading of learning systems in artificial intelligence. A good general discussion of many of the older systems proposed can be found in Cohen and Feigenbaum.<sup>5</sup> Since this thesis deals mainly with induction in connection with problems of constructing botanical keys, the systems which produce decision trees are of particular interest. Prominent amongst these is Hunt's Concept Learning algorithm (CLS).<sup>6</sup> Durkin comments

---

<sup>1</sup>Michalski, Ryszard S., Chilausky, R. L., 'Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis', in *International Journal of Policy Analysis and Information Systems*, Vol. 4, No. 2, June 1980, p. 157.

<sup>2</sup>*Idem.*, p. 151.

<sup>3</sup>Williams, Graham J., *Some Experiments in Decision Tree Induction*, The Australian Computer Journal, Vol. 16, No. 2, May 1987, pps. 84 - 91.

<sup>4</sup>Collier, P.A., "Computer Key Generation from Quantitative Data", (unpublished manuscript). "*1<sup>st</sup> Class*" in this document refers to *1<sup>st</sup> Class* Version 3.

<sup>5</sup>Cohen & Feigenbaum, pps. 323-511.

<sup>6</sup>*Idem.*, p.406; also Quinlan, J. Ross, *Induction of Decision Trees*, Technical Report 85.6, School of Computing Sciences, New South Wales Institute of Technology, 1985, p. 7.

that a descendant of CLS, ID3, is "the most popular [induction algorithm] used today in the design of Expert Systems".<sup>1</sup> As previously commented in section 2.1.3, in ID3 Shannon's entropy function is combined with a sampling similar to that advocated by Edgington to obtain algorithms of significant computational efficiency. The size of the training set will affect the rules induced by the system.<sup>2</sup> A decision tree based on dichotomous and polychotomous<sup>3</sup> decisions can be produced. ID3 has the major advantage of being particularly computationally efficient, although it does show a bias towards characteristics with multi-valued attributes.<sup>4</sup> Cohen and Feigenbaum comment that many possible trees can be drawn from the same data, and that it is difficult to compare them.<sup>5</sup> Further, 'it is difficult for people to understand the learned concept when it is expressed as a large decision tree'.<sup>6</sup> Michie and Quinlan concur with Cohen and Feigenbaum's opinion.<sup>7</sup> However, despite problems in using the resulting decision tree, inductive classification via ID3 and CLS have produced useful results, as will be discussed later.

## 2.2 Obtaining Rules for use in Expert systems

Section 2.2.1 notes that the inductive learning process is potentially useful in extracting structured knowledge from

---

<sup>1</sup>Durkin, *AI Expert*, April 1992, p. 48. For more details of the ID3 algorithm, see Cohen & Feigenbaum, p 406; also Quinlan, J. Ross, 'Learning Efficient Classification Procedures and their Application to Chess End Games', in Michalski, Ryszard S., Carbonell, Jaime G., & Mitchell, Tom M. (Eds.), *Machine Learning, An Artificial Intelligence Approach*, Tioga Publishing Company, Palo Alto, 1983, pps. 463-482.

<sup>2</sup>Edgington, pps. 152-161, also Quinlan, 'Learning efficient classification procedures and their application to chess end games', pps. 463 - 482.

<sup>3</sup>Brown comments: 'To finally lay "polychotomous" to rest I looked it up in the Shorter Oxford Dictionary. It does not exist. The correct word is polytomous. "Polychotomous" is apparently a misapplied generalization of dichotomous.' See: Brown, P., Discussion of paper: Payne, R. W. and Preece, D. A., 'Identification Keys and Diagnostic Tables: A Review', *Journal of the Royal Statistical Society, Series A*, Royal Statistical Society, London, 1980, p. 282. Despite Brown's comment, it seems to have remained customary to use the word "polychotomous" in botanical papers, hence "polychotomous" is used in this thesis in preference to the possibly preferable "polytomous".

<sup>4</sup>Quinlan, *Induction of Decision Trees*, p. 19. See also the prior discussion on this point referred to in section 2.1.3 of this thesis.

<sup>5</sup>Cohen & Feigenbaum, p 410; also Quinlan, *Induction of Decision Trees*, p. 37.

<sup>6</sup>Cohen & Feigenbaum, p 410.

<sup>7</sup>Quinlan, *Induction of Decision Trees*, p.37. See also Michie, D., 'Current developments in Expert Systems', *Proceedings of the Second Australian Conference on Applications of Expert Systems*, Sydney, 1986.

situations where the expert 'does it by eye' and is not aware of the rules he or she uses when applying his or her expertise. Some cases where this approach has proven useful are noted in section 2.2.2. Some disadvantages of the presently available systems are also noted. Section 2.2.3 discusses problems which can occur with data concerned with living specimens, with particular emphasis on botanical specimens.

### 2.2.1 Collecting the Expertise.

Collecting the expertise needed by an expert system has proven slow and sometimes difficult in practice. Quinlan comments that the average knowledge engineer, when interviewing an expert, can expect to gain only a few rules per day<sup>1</sup>. Since an expert system may require several hundred to several thousand rules, building the knowledge base can be both slow and expensive in terms of both time and money.<sup>2</sup>

Partridge, Modesitt and others have commented on the difficulty some experts find in expressing their expertise in the form of rules, (as has been already discussed).<sup>3</sup>

However it is also possible that an expert may be one of Bee's half to one-third of the general population in Western society who do not reach the level described in Piaget's genetic epistemology as the formal or propositional operations stage. It is quite feasible that in some situations an expert could operate purely by induction, (without any deductive ability), classifying a series of characteristics together with a course of action; (e.g. as in "When you get that noise and it starts to shake, you kick it

---

<sup>1</sup>Quinlan, *Induction of Decision Trees*, p. 2.

<sup>2</sup>Once the rules have been obtained, Mao comments that the database structure in which the rules are held can also benefit from an inductive interpretation; see: Mao, Chengjiang, 'THOUGHT: An Integrated Learning system for Acquiring Knowledge Structure', in Sleeman, Derek and Edwards, Peter, *Machine Learning Proceedings of the Ninth International Workshop*, Morgan Kaufmann Incorporated, 1992, pps. 300-309.

<sup>3</sup>Partridge, D., 'Is Intuitive Expertise Rule Based?', in *Proceedings of the Third International Expert Systems Conference*, Learned Information Ltd., Oxford, 1987, p. 346; also Modesitt, K. L., 'Experts: Human and Otherwise', in *Proceedings of the Third International Expert Systems Conference*, Learned Information Ltd., Oxford, 1987, p. 340; also Collins, H. M., 'Domains in Which Expert Systems Could Succeed', *Third International Expert Systems Conference*, Learned Information Inc., Oxford, 1987, p. 204.

there and the noise goes away and it doesn't shake any more").<sup>1</sup> In this case there is little idea of cause & effect, (what caused it to shake, the absence of a kick?). No deductive reasoning of the type used in production rules may be present. The concepts of abstract deductive reasoning may in fact be beyond the expert's grasp, and hence irrelevant to him or her.

In cases like this when a domain expert has not developed abstract deductive logic, and the knowledge engineer has, then either:-

a) a very perceptive and empathic approach is required by the knowledge engineer,

b) an inductive approach is taken based on records of the expert's past judgements & actions, or

c) (preferably) the domain expert can be "built in" to the expert system's inductive categorisation feedback loop in such a way that no more than inductive concepts are required.

Use of the second and/or third options means that some knowledge stored by Mishkin and Appenzeller's postulated second system of learning may be able to be accessed. This may be achieved by making the key-building process interactive, (a conclusion which Wierzbicki notes is occurring in several other areas).<sup>2</sup> It is suggested that this method may be a way of widening the Feigenbaum bottle-neck.

### 2.2.2 Induction and the Feigenbaum bottle-neck.

As suggested before, it is possible to use inductive classification algorithms to partially overcome this bottle-neck.

Buntine notes that inductive algorithms such as Quinlan's ID3 have been successful in simple frameworks, and that when they

---

<sup>1</sup>Less usually, the knowledge may also be procedural (the expert knows how to do it) but not declarative (the expert can not state how to do it).

<sup>2</sup>Wierzbicki, A., 'Interactive decision analysis and interpretative computer intelligence', in *Interactive Decision Analysis*, Springer-Verlag, Berlin, 1984, p. 3; also Pankhurst, R. J., 'An interactive program for the construction of identification keys', *Taxon* Vol. 37., No. 3, August 1988, pps. 747 - 755.



do succeed, considerable gains result.<sup>1</sup> An example of the type of gain is given when Stirling and Buntine discuss the application of inductive techniques in an industrial setting; (routing work through stations in a steel mill).<sup>2</sup> In this use, the rules found through simulated induction were referred back to the expert, and several errors were found in the training set. Stirling also found that several rules were discovered (using Quinlan's C4<sup>3</sup> algorithm) that the expert did not know, but acknowledged as correct when he had examined them.<sup>4</sup>

Bratko and Michie refer to industrial, medical and agricultural applications of inductive logic.<sup>5</sup>

Carter and Catlett refer to industrial applications, and report on the applicability of inductive techniques to credit card applications.<sup>6</sup>

Collier reports on classification of botanical data, (specimens in the *Acaena ovina* complex).<sup>7</sup> Some thirty factors were examined, and botanical keys produced using an algorithm employed in the expert system shell *1<sup>st</sup> Class*. The shell uses the ID3 algorithm.<sup>8</sup>

Collier comments that this approach had the following advantages over unassisted evaluation by the researcher:-

- Much of the work of sifting through the thirty factors is done by *1<sup>st</sup> Class*;

---

<sup>1</sup>Buntine, Wray, 'Decision Tree Induction Systems: A Bayesian Analysis', in *Uncertainty in Artificial Analysis*, publisher unknown, Seattle, 10 July 1987, p. 1.

<sup>2</sup>Stirling, David, & Buntine, Wray, *Process Routings in a Steel Mill, a challenging induction problem*, New South Wales Institute of Technology, Broadway, N.S.W., 1987.

<sup>3</sup>C4 adds selective pruning to ID3 output, see Quinlan, J. Ross, *Simplifying Decision Trees*, Technical Report 87.4, New South Wales Institute of Technology, Sydney, 1987.

<sup>4</sup>Stirling, private communication.

<sup>5</sup>Bratko, Ivan, & Michie, Donald, 'Some comments on rule induction', in *The Knowledge Engineering Review*, Cambridge University Press, Cambridge, Vol. 2, No. 1, March 1987, p. 66.

<sup>6</sup>Carter, Chris and Catlett, Jason, 'Credit Assessment using Machine Learning', *IEEE Expert*, Fall 1987, pps. 71-79.

<sup>7</sup>Collier, P. A., *Inductive Inference for Botanical Keys*, R87-1, Information Science Department, University of Tasmania, Hobart, 1987, p 5.

<sup>8</sup>Collier, private communication; ID3 is documented in Quinlan, J. Ross, *Induction of Decision Trees*.

- Every item of data is represented in the key produced by *1<sup>st</sup> Class*, (hence even outlying values are represented, which is useful in the case of multi-modal distributions).
- The decision tree (botanical key) may be produced by *1<sup>st</sup> Class* in a layout suitable for publication, theoretically not needing re-drafting work by the researcher;

The disadvantages of using *1<sup>st</sup> Class* are:-

- Every example of data is represented in the key produced by *1<sup>st</sup> Class*, (which can be a problem when distributions overlap, as multiple end nodes are found for the same taxon).<sup>1</sup>
- A slight change in the data results in different keys, some being noticeably more elegant than others.<sup>2</sup>
- There is no way to give priority to preferred factors (e.g. those most obvious to the naked eye, those most easily measured, or those available for observation during any season).<sup>3</sup>
- An expert still has to examine the output from *1<sup>st</sup> Class*, and may have to "prune the tree" to make it acceptable as a tool for general identification.<sup>4</sup> (In practice this will often cancel out the third advantage).

Summarising, the fact that *1<sup>st</sup> Class* does the majority of the work in selecting questions and printing them out in acceptable

---

<sup>1</sup>Later implementations than *1<sup>st</sup> Class* include Quinlan's work on automatic pruning of the decision tree to lessen this problem.

<sup>2</sup>Quinlan, *Induction of Decision Trees*, page 139. Also Saxena presents an algorithm which is claimed to evaluate 'which among a given set of alternative [data] representations of a problem is best suited for learning'; see Saxena, Sharad, 'Evaluating Alternative Instance Representation', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishing Inc., San Mateo, U.S.A., 1989, pps. 465-468.

<sup>3</sup>Some factors may be disabled, if the researcher judges them to be of no interest. Also some implementations allow biasing the data to influence the final format of the tree.

<sup>4</sup>Collier, P. A., *Inductive Inference for Botanical Keys*, in Proceedings of the Third Australian Conference on Applications of Expert Systems, The New South Wales Institute of Technology, Sydney, 1987. As noted in a previous footnote, automatic pruning is available in some implementations; in the case of botanic keys, the expert would have to compare the pruned and unpruned trees to check that no required information or taxa have been removed by the pruning process.

format, are both strong plus factors for an overworked researcher.

The fact that all data examples are represented in the output key is both an advantage and a disadvantage. It is an advantage in that exceptional values are not ignored (however one is left with the question of how representative the sample is of the general population; multi-modal distributions seem far less usual than uni-modal distributions in nature). It can be a distinct disadvantage when two distributions are similar, as multiple end nodes can result; for example consider the data postulated in Table 1, (where leaf length is the only available characteristic).

<u>Species</u>	<u>Leaf Length</u>
A	0.9
B	1.3
A	1.9
B	2.3
A	2.4
B	2.8
A	3.4
B	3.8

Table 1 — A possible leaf length distribution

*1<sup>st</sup> Class* would produce eight separate conclusions in a key for this example. A human decision-maker would probably suspect that species A & B had similar leaf lengths, with A being somewhat shorter than B on average (perhaps similar in form to the situation represented in Figure 10), and that the eight conclusions were more a result of an unusual choice of specimens for measurement, rather than several different variations of the same species.

This effect was noticeable in the first experiment Collier reports, which involved submitting measurements of 50 examples of botanical specimens to *1<sup>st</sup> Class*. This process led to a tree with 39 end nodes for 11 different taxa.<sup>1</sup> These could only be reduced by the expert 'pruning' the tree subsequent to its production, or by selectively omitting 'errant' examples from the training set before submitting the data to *1<sup>st</sup> Class*, (although the

---

<sup>1</sup>A similar problem was found when an attempt was made to obtain a no-flower no-fruit key for the *Acaena ovina* complex, see Fig. 24 of this thesis, plus the discussion surrounding this Figure.

latter course of action leaves one with the uneasy feeling that one is 'fiddling the data', a somewhat unscientific behaviour).<sup>1</sup>

Regarding disadvantage c), Collier noted that some of the characteristics chosen by 1<sup>st</sup> Class as splitting characteristics were unusual, 'some strange decisions also appear such as the number of leaflet serrations, and the leaflet length'.<sup>2</sup> Such unusual decisions, chosen on the basis of a mathematically optimal decision selection system, in this case conflicts with the expert's 'gut feeling' as to what was reasonable, and would tend to make the decision tree less understandable and acceptable. Partridge stresses that it is of paramount importance that the rules be understandable.<sup>3</sup> If they are not, there will be difficulty in the expert "trimming" them, or of "fine tuning" the resulting expert system.

Another general disadvantage noted by Bloomfield is that current tools do not allow combinations of the characteristics, each characteristic being treated in isolation.<sup>4</sup> This applies in the case considered by Collier, who notes a key to identify *Acaena* constructed by Orchard includes combinations of characteristics.<sup>5</sup> 1<sup>st</sup> Class does not detect this sort of relationship.

Rendell identifies a fundamental problem with the general class of ID3-like algorithms, when he comments:

In empirical learning, systems for **selective induction** (SI) such as ID3 ... partition instance space into regions of locally invariant or similar class membership values. Recent theory (for any algorithm in the Boolean case) and experiment (for typical systems in the probabilistic case) have shown that

---

<sup>1</sup>Note the publicity recently given to the case of Dr. William McBride.

<sup>2</sup>Collier, p 6. This is an example of the unease felt by some taxonomists when examining the results of automatic key generation processes. For further discussion on this point, see section 2.2.3 of this thesis.

<sup>3</sup>Partridge, p. 346.

<sup>4</sup>Bloomfield, pps. 58-59. However Bloomfield may not have been familiar with the work of Hill; see Evans, D. F., Hill, M. O. & Ward, S. D., *A dichotomous key to British submontane vegetation*, Occasional Paper No. 1, Institute for Terrestrial Ecology, Bangor, North Wales, 1977.

<sup>5</sup>Collier, private communication. For Orchard's key see Orchard, A. E., 'Revision of the *Acaena Ovina* A. Cunn. (Rosaceae) Complex in Australia', Trans. Roy. Soc. S. Aust. (1969), Vol. 93, pps. 91-109. This *Acaena* key is reproduced as Figure 26 in this thesis.

methods of selective induction founder if the membership function has too many disjuncts or *peaks*. ... SI behaviour becomes intolerable when the peaks number in the hundreds, yet important problems (such as protein folding) exhibit millions of peaks.<sup>1</sup>

Rendell goes on to comment that 'transforming the instance space to diminish peaks is one purpose of constructive induction'<sup>2</sup> and goes on to examine the effect of combining selective and constructive induction.

Other methodologies have also been employed, and Matheus reports progress on comparing six systems used for induction.<sup>3</sup> Also induction need not be restricted to just facts and numbers; Bala et. al report on a system which can inductively recognise images.<sup>4</sup>

However whichever methodologies are employed, Dietterich notes as a:-

fact that inductive learning methods are fundamentally limited to learning only a small fraction of possible hypotheses [and that this] has many implications. ... it means that there are no general purpose learning methods that can learn any concept (from a sample of reasonable size). Instead, different classes of learning problems may call for different learning algorithms.<sup>5</sup>

---

<sup>1</sup>Rendell, Larry, 'Comparing Systems and Analysing Functions to Improve Constructive Induction', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishing Inc., San Mateo, U.S.A., 1989, p. 461. The emphasis is as in the original document.

<sup>2</sup>Ibid..

<sup>3</sup>Matheus, Christopher, 'A Constructive Induction Framework', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishing Inc., San Mateo, U.S.A., 1989, p. 475.

<sup>4</sup>Bala, Jerry W., Michalski, Ryszard S. and Wnek, Janusz, 'The Principal Axes Method for Constructive Induction', in Sleeman, Derek and Edwards, Peter, *Machine Learning Proceedings of the Ninth International Workshop*, Morgan Kaufmann Incorporated, 1992, pps. 20-29.

<sup>5</sup>Dietterich, Thomas G., 'Limitations on Inductive Learning', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishing Inc., San Mateo, U.S.A., 1989, p. 128. However also see Almuallim, Hussein and Dietterich, Thomas G., 'On Learning More Concepts', in Sleeman, Derek and Edwards, Peter (Eds.), *Machine Learning: Proceedings of the Ninth International Workshop*, Morgan Kaufmann Publishers, San Mateo, 1992, pps. 11-19 where they produce algorithms which 'have much better coverage than the popular ID3 and its relatives', but which 'strike [them] as trivial' because 'coverage analysis alone is not sufficient ... .

In the case of this thesis, where a methodology is required for the problem of constructing keys to be used to aid the identification of botanic specimens, it would seem that even if a methodology is proposed and developed, several competing methodologies should be compared with the proposed methodology. If Dietterich is correct, all the competing methodologies may be, to at least some extent, problem specific.

### 2.2.3 Common Problems with Data of Botanic Origin.

There are particular problems which occur fairly frequently in sets of botanic data.<sup>1</sup> Common problems include:-

1) *The data sets are often (usually?) unable to meet the usual statistical standards required for the data to be accepted as a statistically valid sample representative of the species/taxa in question.*

For data sets to truly represent the population from which they are drawn, they should represent the product of a carefully designed statistically valid sampling methodology applied to the entire population in question. This is rarely possible with botanic populations.<sup>2</sup> Orlóci comments:

A plant community may have an extent far beyond the possibility of complete enumeration. One may have to be satisfied with statistical estimation (rather than exact determination) of the population parameters.<sup>3</sup>

This means that, strictly, any conclusions drawn from consideration of the results of processing that data should be restricted to that data alone. In practice (because the statistically desirable sampling methodology may be either impractical or impossible to implement) this restriction is often ignored and useful results still obtained. It will be noted that the same theoretical restriction (that the results obtained strictly only apply to the data being examined) also applies to the non-

<sup>1</sup>For a fuller discussion from the point of view of the data used in this thesis, see Appendix E of this thesis.

<sup>2</sup>For example, see the problems faced by Evans *et. al.* in: Evans, D. F., Hill, M. O. & Ward, S. D., *A dichotomous key to British submontane vegetation*, Occasional Paper No. 1, Institute for Terrestrial Ecology, Bangor, North Wales, 1977.

<sup>3</sup>Orlóci, László, *Multivariate Analysis in Vegetation Research*, Dr. W. Junk, The Hague, 1978, p. 189.

parametric methodologies discussed in section 3.1.3 of this thesis, and that hence the conclusion may be drawn that these non-parametric methodologies are less restricted in practice when used with data sets of botanic origin than would seem to be the case when they are considered purely from a theoretical viewpoint.

2) *Many of the characteristics to be observed are either not visible all the time, or are difficult to read.*

Very often the most important characteristics used to distinguish species are the flowers and seeds. These occur seasonally. The expert constructing the key must keep in mind the likely users of a key and the relevance of the available characteristics when preparing that key. For example, a flower might uniquely identify a species, but if it is only visible at night for one night of the year<sup>1</sup> it would be of limited use if included in a key intended for year-round use. The flower characteristics could still be included, but other characteristics would also be needed to cover the rest of the year. This may be the reason Erdtman comments 'No taxonomist, however, would endeavour to classify a plant simply on the basis of a single characteristic'.<sup>2</sup> Thus it is of prime importance that the key construction methodology to be proposed be able to handle multiple characteristics per decision point. The use of multiple characteristics also has the serendipitous effect of helping deal with what Pankhurst notes are two sources of errors:

The specimen may belong to a taxon which is not included in the key ... Hence as many details as possible of each taxon should be used in the key, in order that taxa which do not belong are seen to disagree.<sup>3</sup>

and

If there is only one character per lead in the key, then a wrong branch can more easily be taken. ... As a precaution against

---

<sup>1</sup>As for example the flower of *Selenicereus grandiflorus* is; see de Wit, H. C. D., *Plants of the World - The Higher Plants*, Volume 1, Thames and Hudson, London, 1963, p. 200.

<sup>2</sup>Erdtman, Holger, 'The Assessment of Biochemical Techniques in Plant Taxonomy', in Hawkes, J. G. (Ed.), *Chemotaxonomy and Serotaxonomy*, Academic Press, London, 1968, p. 242.

<sup>3</sup>Pankhurst, 1971, p. 358.

errors of this kind, keys which have several characters per lead are preferred, since there is less doubt if several characters agree.<sup>1</sup>

Cain quotes Darwin as agreeing on the importance of multiple characteristics:

The importance, for classification, of trifling characters, mainly depends on their being correlated with several other characters of more or less importance. The value indeed of an aggregate of characters is very evident in natural history. Hence, as has often been remarked, a species may depart from its allies in several characters, both of high physiological importance and of almost universal prevalence, and yet leave us in no doubt where it should be ranked. Hence, also, it has been found a classification founded on any single character, however important that may be, has always failed; for no part of the organisation is invariable constant.<sup>2</sup>

The characters measured may not all be equally suitable for use in a key.

It is often stated that 'good' characters should be used in keys. A good character is one which is both easy and cheap to determine and which has a high probability of being correctly read.<sup>3</sup>

Since this type of information is not usually represented specifically in the set of botanic data, the choice of 'good' characteristics involves the human expert in an exercise of the application of 'background knowledge' or 'common sense', as:

certain decisions can only be made in the light of practical experience, to suit the specific requirements of the ecologist.<sup>4</sup>

---

<sup>1</sup>*Ibid.*

<sup>2</sup>Cain attributes this quotation to Darwin, C., *On the Origin of Species*, Murray, London, 1859, Chapter 13 (not seen); quoted in Cain, A. J., 'The Assessment of New Types of Character in Taxonomy', in Hawkes, J. G. (Ed.), *Chemotaxonomy and Serotaxonomy*, Academic Press, London, 1968, p. 230.

<sup>3</sup>*Ibid.*

<sup>4</sup>Williams, W. T. and Lambert, J. M., 'Multivariate Methods in Plant Ecology', *The Journal of Ecology*, Volume 48, Blackwell Scientific Publications, Oxford, 1960, p. 690.



The simplest way to most easily and transparently achieve this is to include the expert key constructor in the decision loop.

Pankhurst further comments 'If ... biological species present different features at different seasons, then a variety of differently ordered keys should be available'.<sup>1</sup> The usual method of producing alternate keys involves constant editing of the data set. Since the brief of this methodology is to produce a system which would be useable by a botanical specialist who should not have to be particularly computer-literate, it would be preferable if the alternate keys could be produced without the necessity of editing the data. Again, this could be done by use of an interactive methodology for key construction with the human expert "in the loop".

3) *The data sets often (routinely?) omit measurements of characteristics which were not observable at the time the data was collected.*

A corollary of 2) is that it will be fairly common for some characteristics not to be available at any particular time of the year. It may be necessary to make multiple collecting trips. If the collection areas are extensive or remote, this may not be practical.<sup>2</sup> Missing values are sufficiently common for a convention to have arisen concerning them; Pankhurst comments 'The convention in biology is to call a missing value "not coded", abbreviated as NC'.<sup>3</sup> This lack of completeness can cause problems, e.g. Quinlan notes:-

ignoring cases with unknown values of the tested attribute leads to a very inferior performance (a bitter pill to swallow, as this is how ID3 ... handles partitioning!)<sup>4</sup>

Since then variations of the general ID3 approach such as IDL have been presented.<sup>5</sup>

---

<sup>1</sup>Pankhurst, 1970a, p. 148.

<sup>2</sup>For example the genus *Acaena* extends over several continents; see: Humphries, Christopher J. and Parenti, Lynne R., *Cladistic Biogeography*, Clarendon Press, Oxford, 1989, Figure 1.5, p. 6.

<sup>3</sup>Pankhurst 1970a, p. 146.

<sup>4</sup>Quinlan, Ross J., 'Unknown Attribute Values in Induction', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishing Inc., San Mateo, U.S.A., 1989, p. 168.

<sup>5</sup>Van de Velde, Walter, 'Incremental Induction of Topologically Minimal Trees', in Porter, Bruce and Mooney, Raymond (Eds.), *Machine Learning: Proceedings of*

Another problem that can occur with missing data is that two specimens of different species can end up with identical sets of descriptive characteristics. It is important that the methodology used can cope with this situation.<sup>1</sup>

Thus although in theory data sets of botanic origin can always be complete, in practice any methodology proposed for dealing with botanic data must be able to deal with missing data. It would also be useful if the restriction noted in Pankhurst (1970b) that 'at least one character has to be fully scored'<sup>2</sup> did not apply.

4) *Not all of the characteristics observed meet the requirement of some statistical processes that the characteristics to be employed in those processes be statistically independent of each other.*

Any methodology proposed for use with botanic species should be able to either cope with this type of lack of independence between characteristics, or at the very least check for it.<sup>3</sup>

5) *Many of the characteristics observed are inherently qualitative rather than quantitative, and inevitably a significant degree of individual human judgement (bias?) is involved in rendering these qualitative characteristics into the quantitative terms needed for computer-based processing.*

Pankhurst comments:

... the vast majority of botanical keys (and descriptions of botanical taxa generally) are concerned mainly with qualitative rather than quantitative characters. ... In the great majority of

---

*the Seventh International Conference*, Morgan Kaufmann Publishers Inc., San Mateo, 1990, pps. 66-74.

<sup>1</sup>Some commercial implementations (e.g. *1st Class* which is claimed to use ID3) may not produce a key if this situation occurs.

<sup>2</sup>Pankhurst, R.J., 'Key generation by Computer', *Nature*, London, Vol. 227, September 19, 1970b, pps. 1269-1270.

<sup>3</sup>For example, if the individual petals of unguiculate or cruciate gamopetalous corollas are greater in length, they will often also be wider. These two characteristics would not generally be considered statistically independent. This type of lack of statistical independence is not at all unusual in botanic measurements, and is the reason some keys employ ratios rather than direct measurements of characteristics.

key construction situations, statistical methods are inapplicable.<sup>1</sup>

Perhaps for this reason, in another publication Pankhurst comments:

For the time being, it seems likely that most identifications of plants and animals will depend on human observation.<sup>2</sup>

The author of this thesis accepts Pankhurst's assertion that botanic characteristics used in key construction are mainly qualitative. However the author does not agree that this necessarily eliminates mathematically-based methodologies provided these are implemented in such a way as to aid rather than replace the domain expert.

In support of this view, the author submits that the element of human judgement occurs more widely than may at first be thought. Freeling comments:

There does appear to be some evidence that individuals do decide on the basis of a threshold level [28], and Dreyfuss *et al.* [7] claim that about half the population will decide if objects are members of a fuzzy set by assigning them full membership if they exceed some threshold, and the other half will assign them membership functions as suggested by Zadeh. ... If this is so, then we have a good case for using Zadeh's calculus for the group decision analysis.<sup>3</sup>

The evidence that a significant proportion of humanity uses a judgemental threshold would suggest the existence of an essentially non-linear element in the judgement of a significant proportion of the human race. This would further suggest that

---

<sup>1</sup>Pankhurst, R. J., (private communication). This is a particularly interesting opinion, as Pankhurst himself is one of the leading world authorities on the construction of botanical keys by use of computer methodologies. The words underlined in the quotation were underlined in the original communication.

<sup>2</sup>Pankhurst, Richard J., *Practical taxonomic computing*, Cambridge University Press, Cambridge, 1991, p. 10.

<sup>3</sup>Freeling, Anthony N. S., 'Fuzzy Sets and Decision Analysis', *IEEE Transactions on Systems, Man, and Cybernetics*, Volume SMC-10, Number 7, July 1980, p. 343. Reference [7] (not seen) is given as: Dreyfuss, G. R. *et al.*, 'On the psycholinguistic reality of fuzzy sets', in *Functionalism*, Grossman, R. *et al.* (Eds.), Chicago, IL: Univ. Chicago 1975, pps. 135-149. Reference [28] (not seen) is given as: Reason, J. T., 'Motion sickness, some theoretical considerations', *Int. J. of Man-Mach. Studies*, Volume 1, pps. 21-38, 1969. For the sake of completeness, these references are included in the bibliography as 'not seen'.

any mathematically-based model assuming a continuous distribution would, at best, be an approximate fit to reality. If the model is not a good fit, theorists could suggest that there is another component involved. Kandel and Byatt comment:

The emerging consensus among decision theorists is a view of probability that frankly admits a *subjective component*. It takes into account that there is an element of human judgement even in the seemingly most objective procedures for determining quantitative probabilities, and it does not require that there be only one correct value unless the evidence logically entails it. The essence of this subjective or personal view is that probability is intimately related to human decision making, reflecting a person's degree of belief that the event in question will actually occur.<sup>1</sup>

If, as Kandel and Byatt assert, the element of human judgement (inherently subjective) occurs widely, then admitting its presence in the key construction process would not seem to be a coherent reason for eliminating supposedly objective methodologies merely because an admittedly subjective process is traditionally (and perhaps necessarily) inherent in some of the overall process of botanic key construction.<sup>2</sup>

If the description of the characteristics does include a subjective component (however small) the inclusion of a human expert in the key construction loop would be most desirable. A human expert could appreciate the full meaning of the characteristic description and take much better account of any non-linearities occurring as a result of any subjective or non-linear components introduced into the data-gathering process by the actions of normal human judgement than could any automatic

---

<sup>1</sup>Kandel, Abraham and Byatt, William J., 'Fuzzy Sets, Fuzzy Algebra, and Fuzzy Statistics', *Proceedings of the I.E.E.E.*, Volume 66, No. 12, December 1978, p. 1624. The italics were in the original article.

<sup>2</sup> The author takes the view that the Selecta-key process to be proposed can still legitimately include statistical/mathematical processes, and can (because of the way it presents the key constructor with options ordered in terms of a mathematically valid measure of separation strength for both normal and non-normal distributions) still be of significant assistance to the key constructor during the construction of a botanic key. The author also does not agree with Mayr, who is quoted by Sokal and Sneath as stating 'in the hands of our less gifted [taxonomist] colleagues even the best computer will produce absurd results', Mayr, E. (1959) quoted in Sokal, Robert R. and Sneath, Peter H. A., *Principles of Numerical Taxonomy*, W. H. Freeman and Company, San Francisco, 1963, p. 271.

decision-making process based algorithmically on an estimation process using a continuous function. By contrast, an automatic process which makes the assumption that the data is objectively described can only work with the data as presented; in terms of A.I., it has no access to the largely subjective type of "common sense" which acts as a background to the choices of human botanists and biologists in this area.<sup>1</sup> Considerations such as these may be part of the reason that Pankhurst comments:

Batch mode key-construction programs have been in use for as long as twenty years, but have not found universal acceptance. Evidence has accumulated that keys produced by batch methods are still regarded as being less than ideal. ... This would be true for any computer-constructed key. ... [The computer-constructed key] is not exactly the kind of key which an expert would have chosen to write.<sup>2</sup>

By "batch mode" Pankhurst is referring to keys produced solely by computer, without human intervention during the key-construction process. Pankhurst continues:

Taxonomic experts prefer to make subjective choices of characters at every stage ... The discussion attached to the review of Payne and Preece (1980) shows that taxonomists, mathematicians and computer programmers differ on this point.<sup>3</sup>

Pankhurst then makes an important point which is vital to the approach taken in this thesis:

The purpose of an interactive key-constructing program is therefore not to increase mathematical refinement in the

---

<sup>1</sup>If one's only interest is to improve the theoretical computational efficiency of an algorithm, then these arguments are not a factor, as one can work in a defined and bounded world where a mathematical model's relevance to "reality" is not an issue, and one can use a standard set of "benchmark" data to test one's algorithm against other approaches without having to worry (or even consider) what the data actually means to the real world. However anyone dealing meaningfully and usefully with the real world can not afford the very considerable luxury of the assumptions implicit in this kind of approach.

<sup>2</sup>Pankhurst, Richard J., *Practical taxonomic computing*, Cambridge University Press, Cambridge, 1991, p. 132.

<sup>3</sup>*Ibid.* The reference to Payne *et al.* is: Payne, R. W. and Preece, D. A., 'Identification Keys and Diagnostic Tables: a Review', *Journal of the Royal Statistical Society, Series A*, Volume 143, 1980, pps. 253-292.

algorithms but to increase the participation by the taxonomic expert.<sup>1</sup>

6) *Many botanic characteristics can be highly variable over time.*

Pankhurst comments:

Care is needed in the use of characters which are known to be variable, because they may cause uncertainty in identification. ... If sufficient constant characters are available, the variable characters can be ignored.<sup>2</sup>

If there are not sufficient constant characteristics, variable characteristics must be used, but:

When characters are variable, the different values can be assigned different probabilities of occurrence. Again, in biological cases, these probabilities are not often measured, since they are not constant.<sup>3</sup>

Knowledge of the likelihood that a characteristic would be constant or variable is part of the expert's background knowledge, and unless the data collection process is unusual in that it a longitudinal one (carried out at constant intervals over a

---

<sup>1</sup>*Ibid.*

<sup>2</sup>Pankhurst, R. J., 'A computer program for generating diagnostic keys', *The Computer Journal*, Volume 13, No. 2, May 1970a, p. 148.

<sup>3</sup>*Ibid.*, p. 146. Note that the probabilities mentioned here are not the type of probability usually associated with this type of key (i.e. the probability that the observed characteristic is associated with a particular species rather than another species) but is the probability that the particular characteristic that is associated with a taxa or species will be able to be observed at all (due to factors such as seasonal variation). As an example, consider a key which might be used to help a novice distinguish between *Cydonia oblonga* and *Fraxinus raywoodii*. Both are deciduous, but there is a period when the latter carries leaves that are a very characteristic and unusually uniform russet colour, while the leaves of the former are still green. If the data being considered was collected during this period, an automatic key generation program would be very likely to choose this characteristic as a separating decision; the difference would be (for data of botanic origin) unusually clear-cut. However a human expert would know that this is not a permanent state of affairs; it only lasts for a period of a few weeks. The exact period varies. Gale force winds may limit this period to a few days (a probability of perhaps 1%) or ideal conditions may extend this period to several weeks (6-8%?). The exact probability is not known, it varies. However an expert would know that the probability would be safely under (say) 10%; this would make the characteristic variable, and not a good first choice for a decision node to be used in the construction of an identification key. It could, however, be a useful subsidiary or additional condition at a decision node. In essence, automatic key generation algorithms lack common sense; and the botanic area is a difficult area for key generation algorithms to work. Including the expert in the loop, if this is possible, provides this element in the key-making process.

period of years) it is unlikely that this type of background knowledge (common sense?) will be directly represented in the data. This again suggests that, for the keys to be useful in practice, it would be wise to find some method which would allow the inclusion of the domain expert in the process of choosing the characteristics to be represented at the nodes during the key generation process.

7) *It cannot be automatically assumed that the form of the distribution of the characteristics observed is Gaussian.*

Although past experience has shown that many of the numerable characteristics used in botanic key construction are Gaussian in form, this cannot be assumed. The key generation process should include some methodology for handling non-parametric data.

8) *It cannot be assumed that the preferable minimum number of observations of characteristics which would allow testing the form of the characteristic's data distribution is obtainable in practice.*

Even if the examples in the data of the characteristic's measurement are drawn from a distribution of Gaussian form, they may not be sufficient in number to enable reliable testing of the null hypothesis that there is no difference between the distribution of the data and a gaussian or normal distribution. As an example, see the discussion of the numbers of specimens and characteristics available in the case of *Acaena echinata* var. *robusta* and *Acaena echinata* var. *protenta* in section 3.1.3.2.1 of this thesis. This is an example of the type of problem which Williams, Dale and Macnaughten-Smith note can appear in:

ecology, where few species may be present and some of these species may be rare; a similar difficulty may arise in such human sciences as psychology, sociology or criminology.<sup>1</sup>

Consideration of this point reinforces the conclusion of the discussion of point 7), that some means of handling non-

---

<sup>1</sup>Williams, W. T., Dale, M. B. and Macnaughton-Smith, P., 'An Objective Method of Weighting in Similarity Analysis', *Nature*, January 25, 1964, p. 426.

parametric distributions should be included in any proposed methodology.

*In summary*, it will be important that any proposed methodology be able to deal with the peculiarities of botanic data at least as well, and preferable better, than competing methodologies.<sup>1</sup>

What is required is a methodology that produces a superior botanic key. It is evident from the forgoing discussion that one of the main problems of many existing methodologies is that they are automatic. There is little (if any) chance in some of these methodologies for the expert to use his or her background knowledge and common sense to influence difficult decisions which 'may be decided from outside the data, using [the expert's] knowledge of the field'<sup>2</sup> to influence the choice of characteristics at a node or splitting point.

The second desirable attribute of a suitable methodology would be for the system to present the expert with some measure of the strength of the alternatives, to assist in the choice of appropriate splitting characteristics. It would be preferable that this measure of strength was statistically valid.<sup>3</sup>

Inclusion of these elements alone would make such a methodology significantly more suited to practical botanical key construction than many of the currently available methodologies. Pankhurst commented 'artificial intelligence work has also been concerned with the construction of decision trees, equivalent to keys, although the application to the biological sciences appears to have been overlooked'.<sup>4</sup> Part of the brief of this thesis is to make sure the problems that are peculiar to the biological sciences are addressed in a process aimed at assisting the expert and hence easing the task of construction of keys which are both practical and useful.

---

<sup>1</sup>While this thesis is mainly concerned with data of botanic origin, it should be noted that similar types of problems occur in many data sets where the data is drawn from observations of living subjects, e.g. biology & psychology.

<sup>2</sup>Williams, W. T., Dale, M. B. and Macnaughton-Smith, P., 'An Objective Method of Weighting in Similarity Analysis', *Nature*, January 25, 1964, p. 426.

<sup>3</sup>To avoid splits which are selected on the basis of statistically inadequate data. This can be a problem with some of the presently used methodologies, e.g. see Figure 24 of this thesis.

<sup>4</sup>Pankhurst, *The Computer Journal*, p. 147.



## 2.3 'Selecta-key' Specification.

What seems to be required was a system which allowed the construction of simple, understandable hierarchical key or tree diagrams from botanic data sets which contained either parametric or non-parametric data with missing values, putative outliers and possibly correlated characteristics. It was judged important that the expert was to be involved in the interactive construction of the key at the node level. The node split data calculated by the program should assist the expert by indicating to him or her whether any splits were reasonable at that node; assuming that some were reasonable it should indicate the 'best' characteristics to use for a split at that node in ranked order; it should indicate the strength of the alternative splitting criteria; and whether multiple characteristics could be used at that node of the key or tree. It should be able to handle both dichotomous and polychotomous splits. The information concerning the individual node splitting points should preferably be presented to the expert in a form that could be understood by a person who has not reached Piaget's formal or propositional stage.

This was taken as the specification for the 'Selecta-key' program, which, while being of more general application, has been tested primarily with botanical data and used to construct keys for species identification.

# A STATISTICAL APPROACH TO INDUCTIVE CATEGORISATION

This chapter will examine the application of statistical methodologies which could be useful in constructing botanical keys.<sup>1</sup> Section 3.1 provides an introduction to different statistical approaches, including comments on when it is appropriate to use parametric and non-parametric methodologies. These methodologies are considered in the light of the use of a single characteristic per key decision. Section 3.2 looks at the use of multiple characteristics per key decision, and the protection this can provide against the type of anomalous variation which can occur within a species.<sup>2</sup> Section 3.3 looks briefly at a simplified offshoot of the methodology discussed in this chapter, named the 'voting' method. Section 3.4 provides a brief summary of the use of statistical methodologies in the production of keys used for the identification of species.

## 3.1 Key decisions using a single characteristic

Section 3.1.1 provides a brief introduction to different statistical approaches, including comments on when it is appropriate to use parametric and non-parametric methodologies. Section 3.1.2 considers theory relating to key construction from data collections for which a parametric approach is considered appropriate. Section 3.1.3 similarly considers the case of data collections for which a non-parametric approach is considered appropriate. These methodologies only consider the use of a single characteristic per key decision.

---

<sup>1</sup> Some of the theory presented in this chapter previously appeared in Collier, P.A. and Faulkner, E.G., *Decision Tree Generation using Statistical Methods & a comparison with other methods*, Second International Symposium on Artificial Intelligence, Monterrey, Mexico, 1989; and Collier, P.A. and Faulkner, E.G., *Interactive Decision Tree Generation using Statistical Methods*, Australian Joint Artificial Intelligence Conference, Melbourne, 1989.

<sup>2</sup> In effect, the use of multiple characteristics per key decision can provide a facility similar to that provided by the use of error-correcting codes (Hamming, Golay etc.) in Information Theory. This is particularly useful in the case of botanic identification due to the large intra-species variation found amongst botanic specimens.

### 3.1.1 Statistics and Inductive Categorisation

The problem faced by many researchers attempting to produce botanical keys is essentially the same as that faced by Galton, that of inductively categorising a large amount of data in such a way as to make visible some overall pattern. In Collier's case, the data was submitted to the expert system shell *1<sup>st</sup> Class*, which uses an inductive algorithm to divide the data into classes.<sup>1</sup> Galton used statistical methods.<sup>2</sup> Because statistics was developed to handle the type of biological variation one finds in humans, it also handles well the types of biological variation found in plant data.

It seems that what is required is some sort of combination of the statistical methodology's ability to perform data compression, together with an ability to form a decision key which is understandable and acceptable to an expert.

Statistics may be divided into two main areas, parametric and non-parametric statistics. Parametric statistics assumes the data has a particular shape or distribution. Non-parametric statistics makes no such assumption. Tests based on parametric distributions are generally more powerful than tests based on non-parametric statistics.

In the course of this work, several distributions including the Poisson and Weibull distributions were considered. The Poisson distribution's strength of dealing well with infrequent occurrences seemed not appropriate. The Weibull distribution is a very versatile one, being able to describe distributions ranging from exponential, skewed normal, to normal in shape. However

---

<sup>1</sup> *1st Class* uses the traditional *modus ponendo ponens* type of deductive logic;

$X \supset Y$

$X$

$\therefore Y$

This type of logic is discussed in many logic text books, such as Hatcher, William S., *The Logical Foundations of Mathematics*, Pergamon Press, Oxford, 1982, p. 5; also Copi, Irving M., *Introduction to logic*, fifth edition, Macmillan Publishing Company, New York, 1978, pps. 251 - 252.

<sup>2</sup> The statistics used in this thesis uses an essentially non-deductive logic of the general class:

$g \& Q$

So (probably)  $p$ .

This type of logic is discussed in Sellars, Wilfred, *Are there non-deductive logics?*, in Luckenbach, Sidney A., *Probabilities, Problems, and Paradoxes*, Dickenson Publishing Company, Inc., Encino, California, 1972, pps. 290-307.

since the samples which will be examined in this thesis are uniformly selections from a much larger sample, the central limit theorem makes the presumption of a normal distribution seem much more reasonable. Kohler states:

As long as we take random samples that are sufficiently large absolutely ( $n \geq 30$ ) but that are fairly small relative to population size ( $n \leq .05N$ ) the [central limit] theorem allows us to infer population parameters from sample statistics without knowing the shape of the population distribution (which is precisely the type of knowledge that is often unavailable). In addition ... the theorem can be adapted for use with discrete as well as continuous distributions.<sup>1</sup>

The simplicity of the combination of the Normal (for large normal samples where Kohler's limits apply), t (for small normal samples where there are less than 30 specimens),<sup>2</sup> and non-parametric (for distributions which failed a test for normality) appears preferable.

In the following discussion, issues relating to inductive categorisation will be discussed, firstly in relation to parametric distributions, (with both large and small sample tests discussed separately), and secondly in relation to non-parametric distributions. The discussion will be couched mainly in terms of the type of 'species identification' problem examined by Collier. In this case the size of the sample in relation to the total population is rarely a problem, and only the sample size limitation ( $n \geq 30$ ) will be referred to in the following discussion.

### 3.1.2 Tests assuming Parametric Distributions

The distribution of measurements in a data collection sometimes fits a particular mathematical model. This section deals with the case where it is appropriate to assume that the data fits the assumption of a Normal distribution. Section 3.1.2.1

---

<sup>1</sup>Kohler, Heinz, *Statistics for Business and Economics*, Scott, Foresman and Company, London, 1985, p. 300, 301, 312.

<sup>2</sup>Garrett, p. 186; other authorities vary; e.g. Kreyszig suggests more than 20 for confidence limits on the mean, (but later p. 963 suggests above 30), and more than 50 in confidence limits on  $\sigma$  in Kreyszig, Erwin, *Advanced Engineering Mathematics*, Fifth Edition, John Wiley and Sons, New York, 1983, p. 952.

considers the case where large samples occur.<sup>1</sup> Section 3.1.2.2 considers the case where small samples occur.

### 3.1.2.1 Large Sample Tests

When the number of items of data is greater than or equal to 30 per group, and an assumption that the null hypothesis "that there is no difference between the distribution of the data being considered, and the expected distribution for a similar set of data drawn from a normal distribution" can not be rejected at an appropriate level of confidence (e.g. 5%) then it can be considered appropriate to use the methodologies described in these sections. Section 3.1.2.1.1 provides an introduction to the properties of a Normal curve. Section 3.1.2.1.2 discusses when Normal curves could be considered separate, and section 3.1.2.1.3 goes on to develop this idea in mathematical terms. Section 3.1.2.1.4 discusses where a separation or "splitting" point<sup>2</sup> may be chosen when two distributions are being considered. In many cases however, more than two distributions will be present; section 3.1.2.1.5 discusses the problem of choosing a splitting point in the presence of multiple distributions, and section 3.1.2.1.6 makes recommendations of methods for dealing with this problem. Section 3.1.2.1.7 looks at the effect of type 1 errors on the effective depth of key which may be obtained from a set of data.<sup>3</sup>

#### 3.1.2.1.1 Introduction — Properties of a Normal Curve.

There are several types of parametric distributions. This discussion will be conducted mainly in terms of the 'normal' or 'Gaussian' distribution used by Galton.<sup>4</sup>

Large samples of botanical data often tend towards the normal distribution, and their behaviour can be predicted by results derived from the assumption of normality.<sup>5</sup>

---

<sup>1</sup>Kohler's limit, greater than or equal to 30; see comments on the previous page.

<sup>2</sup>The 'splitting point' would correspond to a decision point on a dendritic key.

<sup>3</sup>A type 1 error occurs when the null hypothesis is rejected, when it is in fact true.

<sup>4</sup>Miller, George A., *Psychology, The Science of Mental Life*, Penguin Books, Harmondsworth, England, 1972, pps. 159-161.

<sup>5</sup>Edgington, Eugene S., *The Distribution-free Approach*, McGraw-Hill Book Company, New York, 1969, p. 73; also Keppel, Geoffrey, *Design & Analysis*, Prentice-Hall Inc., New Jersey, 1973, p. 85.

A normal curve has the form shown in Figure 4.<sup>1</sup>

In the case of the normal distribution, the average value is referred to as the mean ( $\mu$ ), and the spread of observations about that mean is represented a statistic called the standard deviation ( $\sigma$ ).

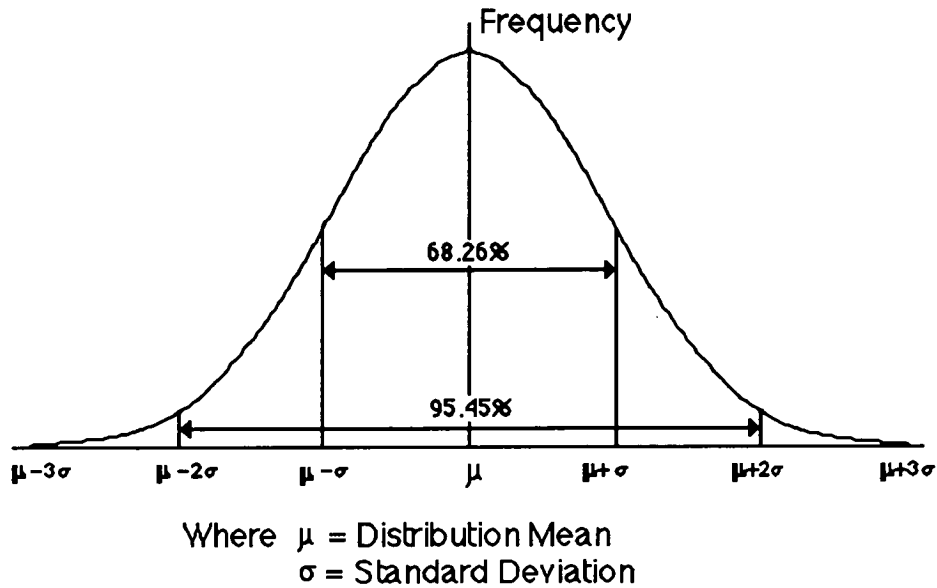


Figure 4 — Normal or Gaussian Probability Curve

Each object or species under study has a series of characteristics which can be observed and (in some cases) measured. Examples of botanical characteristics include leaf length, number of stamens, length of spines on fruit, and so on. The variant of each group of observations of a species characteristic, (if that characteristic is mensurable), can be plotted along the horizontal axis of Figure 4. The set of (e.g.) leaf length observations, if in normal form, can be represented by a number describing the average value of that characteristic (the mean  $\mu$ ), and another describing the spread of observations about that average value (the standard deviation  $\sigma$ ). In this way the essential characteristics of the leaf length observations can be reduced to two numbers, much simplifying subsequent mathematical manipulation.

It should be noted that if the characteristic is categoric, (e.g. the presence or absence of a centre vein in a leaf), then this

<sup>1</sup>Garrett, Henry E., *Statistics in Psychology and Education*, Vakils, Feffer & Simons Pty. Ltd., Bombay, 1967, p. 89.

method is less applicable, but can be used if required by employing the subterfuge of representing each state with a number, and accepting the implied ordering of the categoric states.<sup>1</sup>

The equation of the normal curve catering for  $n$  cases is:<sup>2</sup>

$$f(x) = \frac{n}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

where  $f(x)$  = the ordinate of the curve for a given  $x$ , i.e. the frequency.

Estimates of a sample's mean<sup>3</sup> and standard deviation<sup>4</sup> may be obtained by the following formulæ:

$$\mu = \frac{\sum_{i=1}^n x_i}{n} \quad (2)$$

where  $\mu$  = mean of the  $n$  points  $x_i$

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}} \quad (3)$$

where  $\sigma$  = standard deviation of the  $n$  points  $x_i$

The number of a random cases of the variant  $x$  which fall between the limits  $x=A$  and  $x=B$  are represented by the shaded area in Figure 5, this area also representing the proportion of the total population that have values between  $A$  and  $B$ . If  $n=1$ , this shaded area can also represent the probability that a case would fall between the limits  $A$  and  $B$ , since when  $n=1$  the area under the curve is 1.0.

<sup>1</sup>Note that in the Selecta-key system, these numbers (and the implied ordering that goes with them) may be chosen by the user.

<sup>2</sup>Burr, Irving W., *Engineering Statistics and Quality Control*, McGraw-Hill Book Company, New York, 1953, p. 66; also Hart, Anna, *Knowledge Acquisition for Expert Systems*, Kogan Page, London, 1986, p. 77; also Garrett p. 96.

<sup>3</sup>Ali, A. M., 'Probability - Uncertainty - Simulation', in Jelen, F. C., (Ed.), *Cost and Optimization Engineering*, McGraw-Hill Book Company, New York, 1970, p. 154; also Andreas, Burton G., *Experimental Psychology*, John Wiley and Sons, Inc., New York, 1960, p. 66.

<sup>4</sup>Dhillon, Balbir S., *Quality Control, Reliability, and Engineering Design*, Marcel Dekker, Inc., New York, 1985, p. 81; also Andreas p. 72.

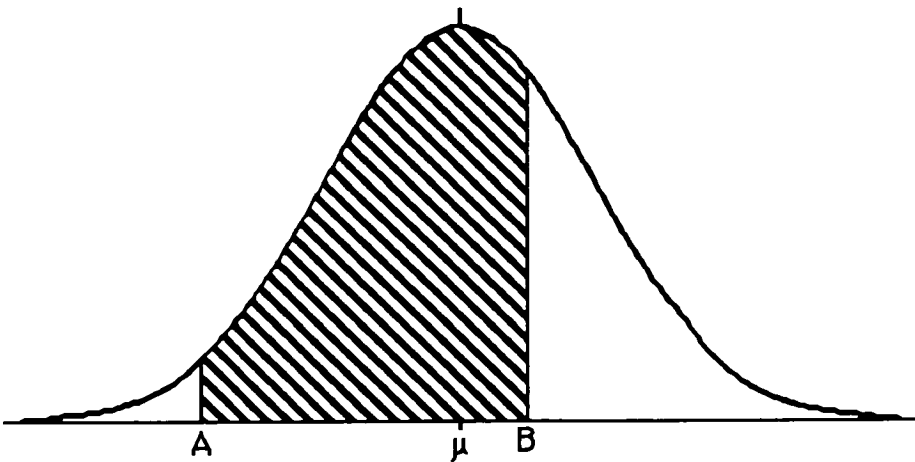


Figure 5 — Area under the normal probability curve

To evaluate the shaded area we employ:<sup>1</sup>

$$p(A \leq x \leq B) = \int_{x=A}^B \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \quad (4)$$

where  $p$  = probability that  $x$  will be in the range  $A \dots B$ .

The area under the curve can be found for any given value of  $\mu$ ,  $\sigma$ ,  $A$  and  $B$ , using an approximate method such as Simpson's rule, or an infinite series. There are also standard tables of the area under this curve in most books on statistics.<sup>2</sup> These give the following areas under the curves.

Interval	Area (probability)
$\mu - 0.67\sigma$ to $\mu + 0.67\sigma$	0.5
$\mu - \sigma$ to $\mu + \sigma$	0.6827
$\mu - 1.96\sigma$ to $\mu + 1.96\sigma$	0.95
$\mu - 2\sigma$ to $\mu + 2\sigma$	0.9545
$\mu - 2.58\sigma$ to $\mu + 2.58\sigma$	0.99
$\mu - 3\sigma$ to $\mu + 3\sigma$	0.9973

Table 2 — Area between confidence limits

<sup>1</sup>Burr, p. 68; also Hart, p. 79.

<sup>2</sup>Hoel, Paul. S., *Introduction to Mathematical Statistics*, John Wiley & Sons, New York, 1954, pps. 315-317; also Knowler, LLOYD A., Howell, John M., Gold, Ben K., Coleman, Edward P., Moan, Obert B., Knowler, William C., *Quality Control by Statistical Methods*, McGraw-Hill Book Company, New York, 1969, pps. 21, 116; also Burr, pps. 404-405; also Garrett, p. 459.



In psychological work, the  $\mu \pm 1.96\sigma$  and  $\mu \pm 2.58\sigma$  limits are referred to respectively as 95% and 99% confidence limits.<sup>1</sup>

Users of engineering quality control similarly use  $\mu \pm 3\sigma$  limits,<sup>2</sup> referring to them as loosely as 99% confidence limits. This would correspond to a type 1 error of .003,<sup>3</sup> however these limits are somewhat rough because:-

many industrial variables are not normally distributed, and since the sample means used in control charts are often based on only 4 or 5 measurements, one could hardly expect the probability of .003 to be very realistic. Three standard deviation control limits are chosen because industrial experience has found them to be especially useful, rather than because they correspond to a desirable probability.<sup>4</sup>

The botanical data we will employ will usually have more than 4 or 5 measurements, but whether experience with the system will find two or three standard deviation limits to be more useful has yet to be established, so it would seem advantageous for any program based on this theory to allow rejection limits to be specified by the user.

#### 3.1.2.1.2 *Distinguishing between two Species.*

The concept of a normal or Gaussian distribution can be useful when applied to species identification. Consider the following case where two distributions have mutually distinct ranges, e.g. the distribution of the diameters of two fruit of widely differing size, such as mature blackcurrants and nectarines (Figure 6).

In this case, a diameter of 5 cms. looks to be well outside the  $\mu \pm 3\sigma$  limits for blackcurrants, and one could reasonably include it

---

<sup>1</sup>Garrett, p. 188.

<sup>2</sup>Bicking, C. A., 'Process Control by Statistical Methods', in Juran, J. M., Gryna Jr., Dr. Frank M., Bingham Jr., R. S., (Eds.), *Quality Control Handbook*, McGraw-Hill Book Company, New York, 1951, p. 23-8; also Dhillon, Balbir S., *Quality Control, Reliability, and Engineering Design*, Marcel Dekker, Inc., New York, 1985, pps. 98, 102; also Knowler et. al., p. 20.

<sup>3</sup>That is, a process operating within tolerance would produce 3 items in 1000 that would be rejected.

<sup>4</sup>Burr, pps. 108-109. Note that (typically in a real-world application such as Engineering) this is an inductively-derived limit.

amongst the nectarines. Similarly an 0.8 cm. diameter fruit could reasonably be included amongst the blackcurrants.

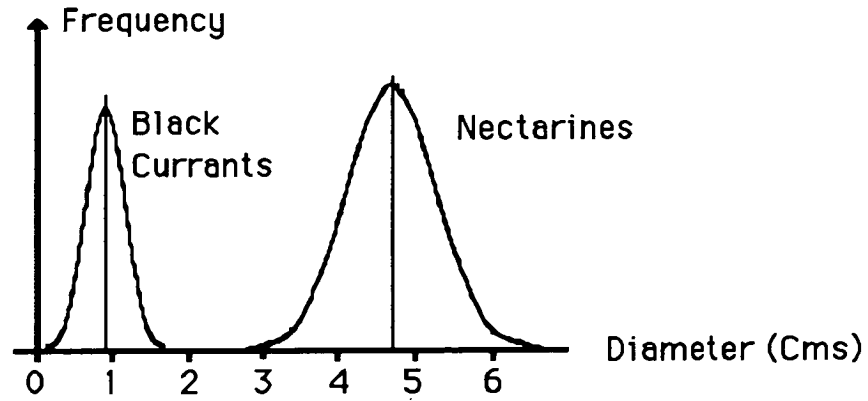


Figure 6 — Distributions exhibiting separation

However it would be preferable to have some more reliable form of distinguishing two distributions than the 'look' of them. Mathematical methods are needed for :-

- a) Establishing that the means of the two distributions are distinct, (according to an acceptable statistical standard); and
- b) If the distributions are distinct, establishing an acceptable dividing line between them.

### 3.1.2.1.3 Are the means of two large-sample distributions different?

A set of observations is a sample or subgroup taken from the whole population of individual values. The mean and standard deviation of the sample will in general not be identical to the corresponding measures of the parent distribution, as it is unlikely the sample exactly represents the distribution from which it was drawn, i.e. sampling errors occur. Assuming an acceptable sampling procedure, the larger the sample, the more likely it is that the sample will be truly representative of the whole distribution. With the same proviso, the larger the sample, the nearer the mean of the sample will be to the mean of the population.<sup>1</sup> The standard deviation of the mean (also called the

<sup>1</sup>Coyne, Anthony M., *Introduction to Inductive Reasoning*, University Press of America, Inc., London, 1984, p. 222.

standard error of the mean) of the sample, is defined by equation 5.<sup>1</sup>

$$\sigma_{\mu_{\alpha}} = \frac{\sigma_{\alpha}}{\sqrt{n_{\alpha}}} \quad (5)$$

where  $\sigma_{\mu_{\alpha}}$  = standard error of the mean  $\mu_{\alpha}$   
 $\sigma_{\alpha}$  = standard deviation of the set of individual observations  $\alpha$   
 $n_{\alpha}$  = number of observations in set  $\alpha$ .

This formula suggests a way of distinguishing between two distributions. The separation between the means of samples taken from the uncorrelated distributions  $\alpha$  and  $\beta$  can be examined by using the formulæ in equations (6) & (7):-<sup>2</sup>

$$\sigma_{\alpha-\beta} = \sqrt{\frac{\sigma_{\alpha}^2}{n_{\alpha}} + \frac{\sigma_{\beta}^2}{n_{\beta}}} \quad (6)$$

where:-  $\sigma_{\alpha-\beta}$  = standard deviation of  $\mu_{\beta} - \mu_{\alpha}$   
 $\sigma_{\alpha}$  = standard deviation of group  $\alpha$ ,  
 $\sigma_{\beta}$  = standard deviation of group  $\beta$   
 $n_{\alpha}$  = number of observations in group  $\alpha$   
 $n_{\beta}$  = number of observations in group  $\beta$ .

and:-

$$CR = \frac{(\mu_{\alpha} - \mu_{\beta})}{\sigma_{\alpha-\beta}} \quad (7)$$

where  $CR$ =critical ratio.<sup>3</sup>

Since  $\alpha$  and  $\beta$  are both large samples, it can be assumed that the distribution of  $CR$  is normal,<sup>4</sup> and if  $CR \geq 1.96$ , there is a 5% chance of a type 1 error in this two-tailed test; i.e. that we reject the null hypothesis<sup>5</sup> (that there is no difference between the two means), when in fact they are from the same population. If  $CR \geq$

<sup>1</sup>Hoel, p. 104; also Moroney, M. J., *Facts from Figures*, Penguin Books Ltd., Harmondsworth, England, 1984, p. 137.

<sup>2</sup>Hoel, p. 109; also Garrett p. 214.

<sup>3</sup>Steel, Robert. G. D. and Torrie, James H., *Principles and Procedures of Statistics, A Biometric Approach*, Second Edition, McGraw-Hill Book Company, Singapore, 1981, p. 95.

<sup>4</sup>Garrett, p. 215.

<sup>5</sup>Andreas, pps. 84 - 85.

2.58, there is a 1% chance that a type 1 error has occurred. However the higher limits also increase the chance of a type 2 error, that of accepting the null hypothesis when it is in fact false.

In preliminary and some psychological work, where it is usually very difficult to control all extraneous variables, a 5% limit is usually regarded as sufficient.<sup>1</sup> In this thesis we suggest a higher limit both because a higher limit has proven more useful in practice and because a method of dealing with non-separated distributions has been developed. Thus in this thesis, if the *CR* is greater than 2 or preferably 3, it is considered reasonable to reject the null hypothesis that there is no difference between the two means and in effect assume that they come from different distributions. If the *CR* is less than 2 or 3, then the data is not sufficient to separate the two sets of observations, and it is reasonably possible that they could come from the same distribution.

Assuming the two distributions are distinct, the next task is to choose a point at which they may reasonably be divided from each other.

#### 3.1.2.1.4 Separation points in large sample parametric distributions.

Consider the distributions shown in Figure 7:-

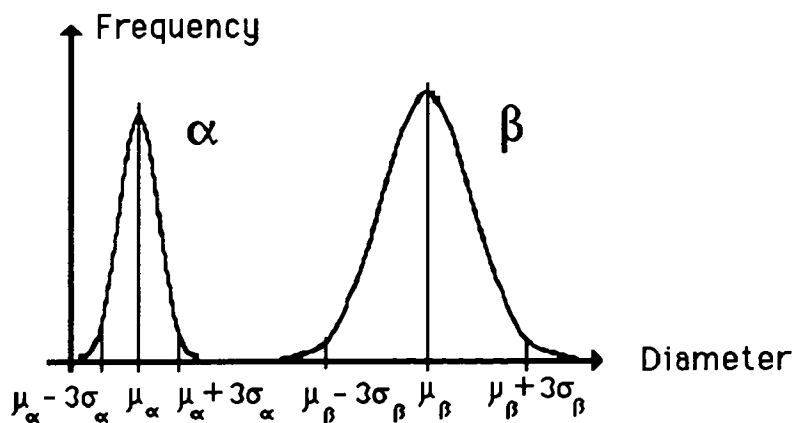


Figure 7 — Distributions exhibiting separation

<sup>1</sup>Garrett, p. 223.

A dividing point could be chosen between  $\alpha$  and  $\beta$ , drawn at a value of the variant  $x$  that we will call  $x_{split}$ .

If the portions of the distributions within their respective  $\mu \pm 3\sigma$  limits do not overlap, (as in Figure 7),  $x_{split}$  could be any point outside these limits between the two distributions. One option would be to choose the point midway between  $\mu_\alpha + 3\sigma_\alpha$  and  $\mu_\beta - 3\sigma_\beta$  limits as an appropriate splitting value, so that  $x_{split}$  would be defined by equation 8:-

$$x_{split} = \frac{\mu_\alpha + 3\sigma_\alpha + \mu_\beta - 3\sigma_\beta}{2} \quad (8)$$

If the two distributions do overlap, as shown in Figure 8, there are several options for  $x_{split}$ .

One option would be to choose the point midway between  $\mu_\alpha$  and  $\mu_\beta$  as an appropriate splitting value, so that  $x_{split}$  would be defined by equation 9:-

$$x_{split} = \frac{\mu_\alpha + \mu_\beta}{2} \quad (9)$$

A second option would be to proportion the space between the means in accord with the standard deviations, as in equation 10.

$$x_{split} = \mu_\alpha + \frac{(\mu_\beta - \mu_\alpha) * \sigma_\alpha}{(\sigma_\alpha + \sigma_\beta)} \quad (10)$$

Another option would be the choice of the point of equal frequency, as represented by Figure 8.<sup>1</sup>

---

<sup>1</sup>Gower prefers this splitting point, see Gower, J. C., 'Relating Classification to Identification', in Pankhurst, R. J., (Ed.), *Biological Identification with Computers*, Systematics Association Special Volume No. 7, Academic Press, London, 1975, p. 255.

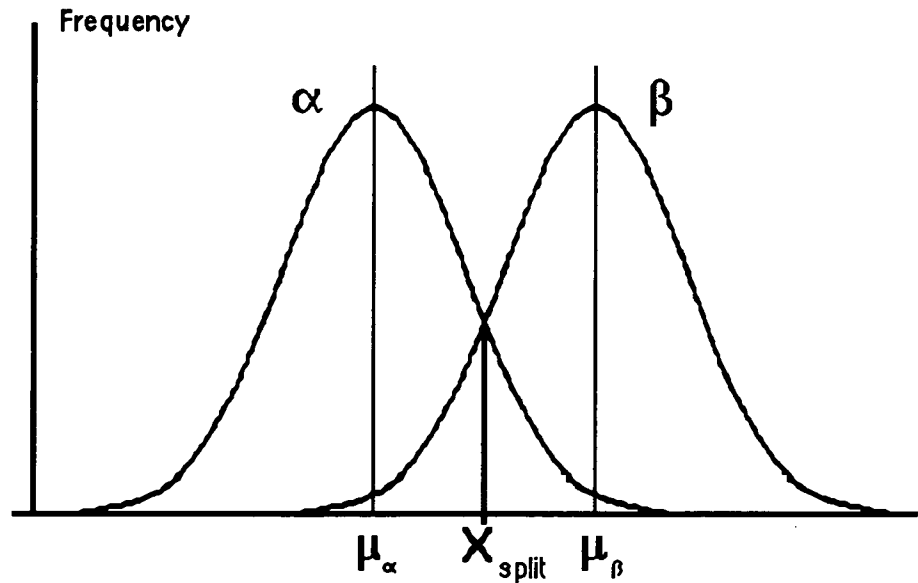


Figure 8 — Overlapping Distributions

Since the frequencies are equal, the following equation for  $x_{split}$  applies:

$$\frac{n_\alpha}{\sigma_\alpha \sqrt{2\pi}} e^{-\frac{(x_{split}-\mu_\alpha)^2}{2\sigma_\alpha^2}} = \frac{n_\beta}{\sigma_\beta \sqrt{2\pi}} e^{-\frac{(x_{split}-\mu_\beta)^2}{2\sigma_\beta^2}} \quad (11)$$

Simplifying this for  $x_{split}$ , one obtains:-

$$x_{split} = \frac{\sigma_\alpha^2 \mu_\beta - \sigma_\beta^2 \mu_\alpha \pm \sigma_\alpha \sigma_\beta \sqrt{(\mu_\alpha - \mu_\beta)^2 + 2(\sigma_\alpha^2 - \sigma_\beta^2) \ln\left(\frac{n_\beta \sigma_\alpha}{n_\alpha \sigma_\beta}\right)}}{\sigma_\alpha^2 - \sigma_\beta^2} \quad (12)$$

The negative sign before the square root is appropriate when  $\mu_\alpha < \mu_\beta$ , otherwise the positive sign is used. (Note that it is appropriate to use both signs if either of  $\sigma_\alpha$  or  $\sigma_\beta$  is very much greater than the other.<sup>1</sup> Note also that in the special case where  $\sigma_\alpha = \sigma_\beta$  and  $n_\beta = n_\alpha$ , that  $x_{split}$  reverts to equation (9).)

A fourth option would be to choose the value of  $x_{split}$  so that it had an equal probability of occurrence in each distribution; i.e. so that the proportion of the shaded areas to the total area under the curve in each of the distributions below are equal, as

<sup>1</sup>See Figure 11.

represented diagrammatically in Figure 9. (For normal distributions whose individual areas equal unity, Figures 8 & 9 would be similar).

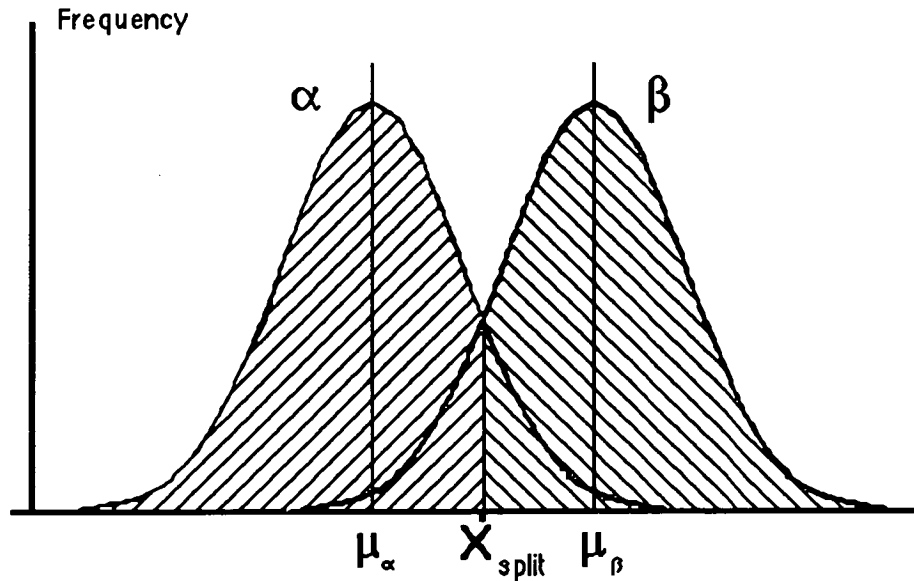


Figure 9 — Overlapping Distributions

The equation for this condition is as follows, (equation 13):-

$$\frac{1}{\sigma_{\alpha}} \int_{x=-\infty}^{x_{split}} e^{\frac{-(x-\mu_{\alpha})^2}{2\sigma_{\alpha}^2}} dx = \frac{1}{\sigma_{\beta}} \int_{x=x_{split}}^{+\infty} e^{\frac{-(x-\mu_{\beta})^2}{2\sigma_{\beta}^2}} dx \quad (13)$$

This is probably best solved for  $x_{split}$  by a numerical method, such as Simpson's rule or an infinite series.

If the value of  $x_{split}$  lies outside the  $\mu \pm 3\sigma$  limits for each distribution, and the sample is expected to contain only data relating to the two items under consideration, it is not critical which of equations 8 to 13 is used to obtain the  $x_{split}$  value, as the samples will be clearly in one category or the other more than 997 times out of 1000. Inside this limit, problems can occur.

Consider the case represented by Table 1.<sup>1</sup> If a normal distribution was assumed, and the mean and standard deviation of the two distributions given in this Table were calculated, the results in Table 3 would be obtained.

<sup>1</sup>See section 2.2.2 of this thesis.

	$\mu$	$\sigma$	S.E.	1 $\sigma$ limits	2 $\sigma$ limits
A	2.15	0.90	1.04	1.25-3.05	0.35-3.95
B	2.55	0.90	1.04	1.65-3.45	0.75-4.35

Table 3 — Mean and standard deviation

If one examines the data from Table 3, it can be seen that the null hypothesis, (that the two distributions are drawn from the same distribution), may not be rejected, as the Critical Ratio = 0.6.<sup>1</sup> Thus this data is not statistically suitable for use in separating species A and B, confirming the postulated suspicion of the researcher<sup>2</sup>

Consider also the case shown in Figure 10.<sup>3</sup>

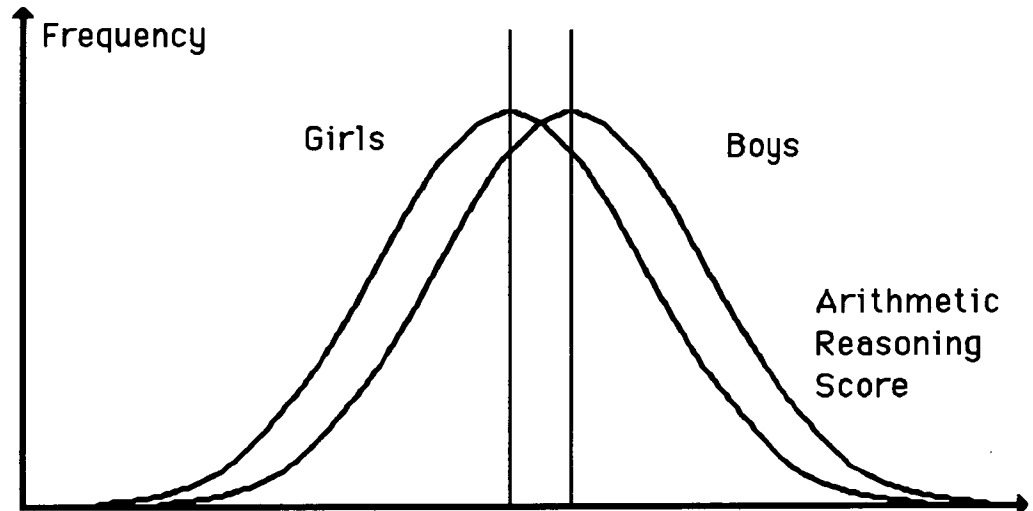


Figure 10 — Scores on an Arithmetic Reasoning Test

The boy's median score was 42, the girl's 32. In this case the use of strict non-overlapping categories and an  $x_{split}$  value of, say, 37 would suggest that any score above 37 would be obtained by a

<sup>1</sup>See Equation 7, section 3.1.2.1.3 of this thesis, plus the surrounding discussion.

<sup>2</sup>However *1st Class* (a commercial program against which this approach was tested) produces a key which would appear (if accepted) to be as valid as any other type of key it produces. This may also constitute an example where a researcher may consider it appropriate to "prune" the decision tree produced by *1st Class*, in the interests of obtaining a useful key. Problems relating to 'pruning' are discussed in section 3.1.2.1.7 of this thesis.

<sup>3</sup>The curves are representative of the ogives given by Garrett in Figure 11 on page 74, but redrawn as a frequency diagram.



boy, and any score below 37 would be obtained by a girl. This is clearly an inaccurate representation of the situation.<sup>1</sup>

1<sup>st</sup> Class also does not do very well in this case. It produces a decision key listing separate conclusions for each non-overlapping range of scores obtained by the girls and boys. In a case such as the one above, a deterministic decision key of this type is of limited use in classifying new data. If a score is obtained and it is desirable to estimate if this score is more likely to be obtained by a boy or a girl, the decision key gives a definite answer based on a deterministic interpretation of the data fed in so far. In this case a statistical interpretation is more appropriate, partly because it uses a model of the data distribution.

Fu comments:

Model-driven learning methods ... are superior in escaping this type of noise because there exist global criteria (which measure the consistency over the instances) for selecting hypotheses generated by the models, and the instances are not considered individually. Since the methods intend to find the most consistent concept descriptions or rules, falsely classified instances will ... be ignored if they are in the minority.<sup>2</sup>

If we employ a statistical model for the form of the data, and the effects of extreme, noisy, or erroneously classified data is minimised, and the number of leaves representing a minimal number of specimens is also minimised.<sup>3</sup>

In the case where a model is employed, a problem may still occur when there is sufficient statistical information to separate the species, but the distance from the relevant statistical means to the splitting point is less than desirable. A diagrammatic

---

<sup>1</sup>This type of problem with overlapping distributions also occurs in many industrial control applications, e.g. see Leitch, Roy & Francis, John, 'Towards Intelligent Control Systems', in Mamdani, Abe & Efstathiou, Janet, (Eds.), *Expert Systems and Optimisation in Process Control*, Gower Technical Press, Aldershot, England, 1986, p. 64.

<sup>2</sup>Li-Min Fu, 'Learning Object-Level and Meta-Level Knowledge in Expert Systems', Technical Report No. STAN-CS-86-1091, Department of Computer Science, Stanford University, 1985, p. 101.

<sup>3</sup>By contrast, ID3 does not employ a model of the data, and is at the mercy of noisy data; however pruning algorithms such as those contained in Quinlan's C4.5 can markedly improve ID3 keys produced from noisy data.

representation of a typical case where this can occur is shown in Figure 11.

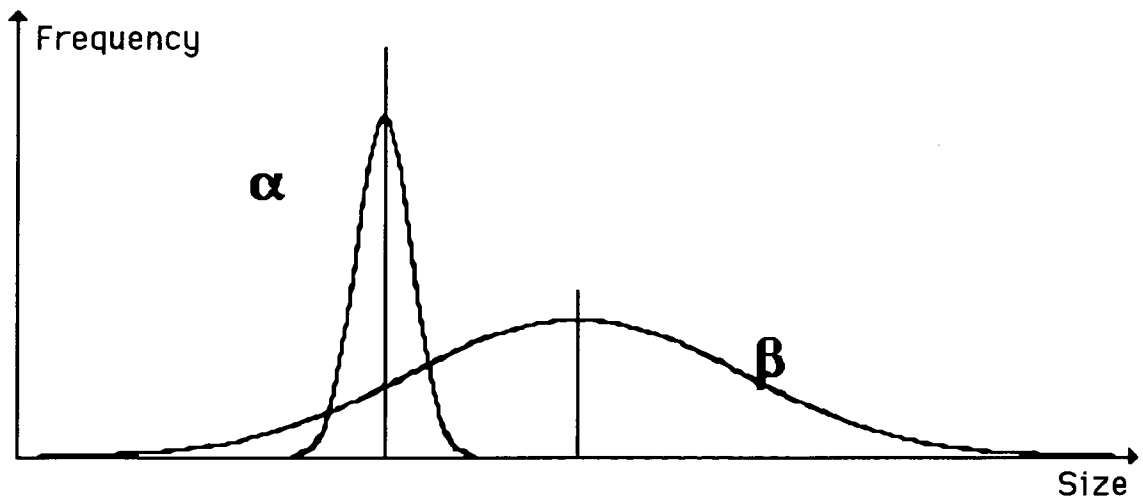


Figure 11 Incompletely separated distributions

In the above case, any single splitting value chosen to separate distributions  $\alpha$  and  $\beta$  would be very likely to be less than a desirable number of standard deviations from the  $\alpha$  and  $\beta$  means, and thus good separation of the two distributions would be unlikely. In this case, a system allowing two splitting points (one on either side of distribution  $\alpha$ ) would be preferable.

#### 3.1.2.1.5 Distinguishing between many large-sample Distributions.

The problem of selecting a splitting point when there are more than two large sample parametric sets of observations is more complex. Consider the example shown in Figure 12 on the next page.

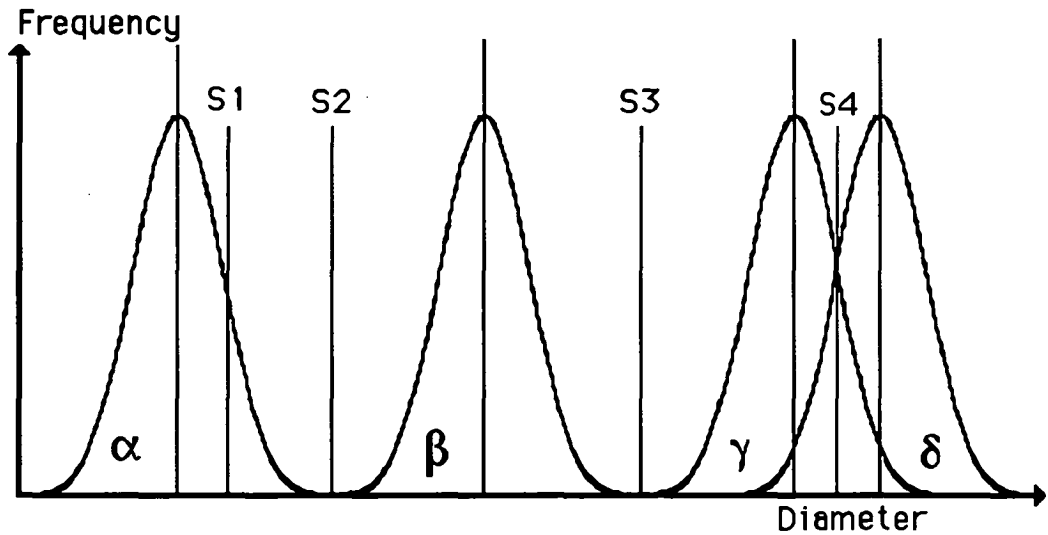


Figure 12 — Multiple Distributions

Here S1, S2, S3 and S4 are possible splitting points. S1 is an arbitrary splitting point. S2 and S3 represent locally optimised splitting points between the distributions on either side of them, chosen by one of the methods outlined previously. S4 is a 'equal frequency' splitting point for distributions  $\gamma$  and  $\delta$ , such as that outlined in Figure 8, and associated equation 12.

S1 would not be an appropriate splitting point, as it does not separate any group from any other.

S3 may be an appropriate splitting point, as it would allow groups  $\alpha$  and  $\beta$  to be separated from groups  $\gamma$  and  $\delta$ . Further work would be needed to separate these groups, although each would need only one further split, producing a 'balanced key' with a depth of two decisions to separate out any of the groups.<sup>1</sup>

S2 would also be a possible splitting point, and would have the advantage of separating out group  $\alpha$  with only one decision. This would be appropriate if the expert estimated that this would be the species required to be most often identified. The disadvantage of choosing this splitting point, however, would be that up to two more decisions would be needed to be made to

<sup>1</sup>Theoretically a 'balanced tree' would be the most economical for use in identifying specimens if all species were equally likely to be presented for identification. If one species was by far the most common, and this could be separated out with one decision, then theoretically this type of tree (an 'unbalanced tree') would be preferable for species identification with this taxa. However this type of optimisation is not usually applicable for botanic species. Pankhurst comments: 'With biological species, numerical measures of the relative abundance of species are not usually available', Pankhurst, 1970, p. 146.

separate out the groups  $\beta$ ,  $\gamma$  and  $\delta$ ; i.e. this initial split would lead to an 'unbalanced decision key'. If the expert is wrong, and specimens of all species are presented for identification equally often, an unbalanced key would lead to more decisions being required (on average) than would be required if the decision key was balanced.

The same problems apply if splitting point S4 is chosen, with the additional disadvantage that groups  $\gamma$  and  $\delta$  would be incompletely separated.

The decisions above assume the key construction methodology only allows dichotomous decisions to be made at each node. If polychotomous decisions are permissible, choosing S2 and S3 would allow  $\alpha$  and  $\beta$  to be separated both from each other and from both  $\gamma$  and  $\delta$ .<sup>1</sup> Depending on the frequency of the presentation of the four species to the key for identification, this could be preferable to the alternatives mentioned above in that only one more node would be needed to separate  $\gamma$  and  $\delta$ .

Once these locally optimised splitting points have been found, the next step is to either:-

- a) present these splitting points to an expert to enable him or her to make a choice based on the expert's knowledge of the appropriateness of the data, splitting point, and potential use of the decision key; or
- b) allow the automatic generation of a decision key, with the choice between the locally optimised splitting points being made on the basis of some pre-determined criteria.<sup>2</sup>

#### *3.1.2.1.6 Splitting Points and Multiple Distributions*

In the case where multiple distributions are present, two possible methods of choosing splitting points present themselves, the 'Grouping' method (section 3.1.2.1.6.1), and the 'Individual Difference' method (section 3.1.2.1.6.2).

---

<sup>1</sup>These are also sometimes referred to loosely as 'n-ary' decisions.

<sup>2</sup>E.g. see the discussion in section 3.1.2.1.6.2 of this thesis.

### 3.1.2.1.6.1 Method 1 — 'Grouping'

In this case, a splitting point is chosen and all the points below this value are assumed for the purpose of this method to belong to one distribution, and all those points above to belong to a second. In this case a test could be used to estimate if the splitting value is a reasonable one, i.e. if the null hypothesis that the two distributions are in fact the same distribution can be rejected with a reasonable level of confidence. Since the samples either side of  $x_{split}$  are made up of a number of distributions summed together, each sample may well be multi-modal, (e.g. Figure 13).

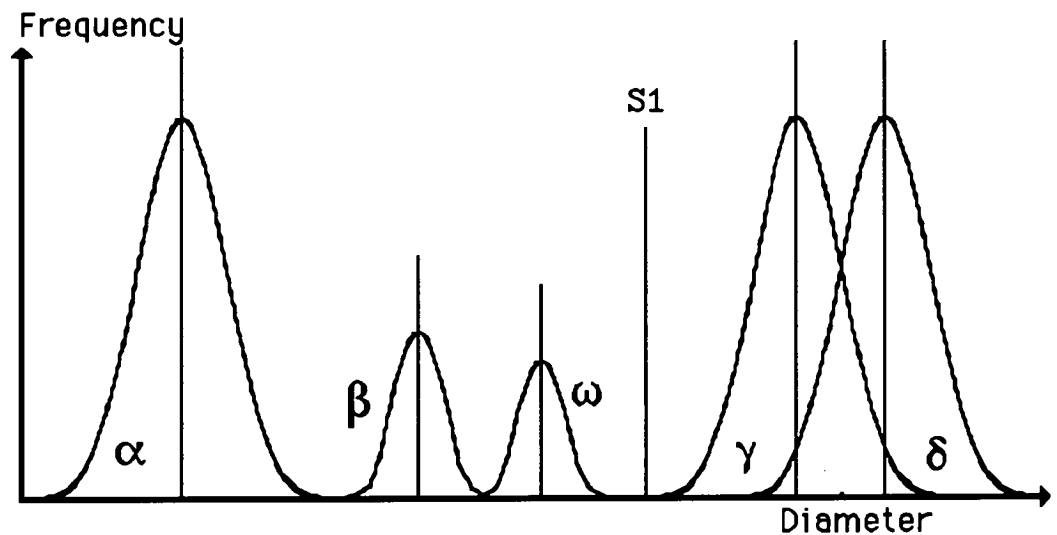


Figure 13 — Multi-modal distributions with splitting point S1.

In this case a non-parametric test such as those discussed later would probably be more appropriate than the type of parametric test specified by equations 5, 6 and 7.<sup>1</sup>

If a grouping methodology is adopted, care must be taken in situations such as those represented by Figure 14, (a situation which can occur often in the collection of botanic data).

<sup>1</sup>Non-parametric tests are discussed in section 3.1.3 of this thesis.

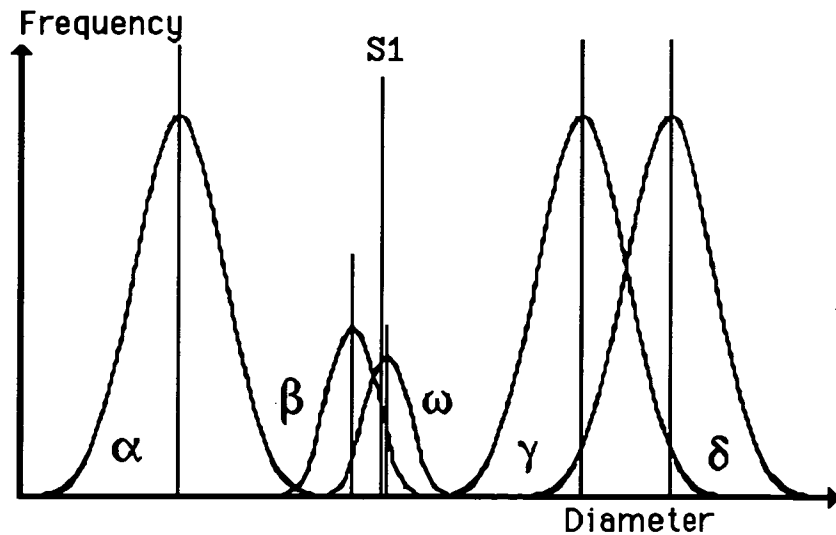


Figure 14 — Multi-modal distributions with small distributions grouped about splitting point S1.

If the distributions  $\alpha$ ,  $\gamma$  and  $\delta$  are much bigger than the distributions  $\beta$  and  $\omega$ , the latter two distributions can be overlooked in the (otherwise justified) choice of a splitting point such as S1 in Figure 14. In this case the S1 splitting point in the decision key would favour the much larger distributions to the detriment of the smaller distributions. While this may be justified in terms of the overall separation of specimens by this particular decision, it poses problems for the construction of an economical key.  $\beta$  and  $\omega$  will each either have to appear at least twice (as a minimum) in the conclusions of that key, or a good proportion of their specimens will be wrongly identified when the key is used to identify specimens. The methodology discussed below can help avoid this trap.

### 3.1.2.1.6.2 Method 2 — 'Individual Difference'

In this case each group of observations is compared individually with another group of observations. This results in  $\sum_{i=1}^n (n-1) \approx O(n^2)$  comparisons for  $n$  observations, but since botanical keys are usually not large, and this calculation only has to be done once, it is anticipated that this will not be too much of a handicap in practice.<sup>1</sup> The results obtained from the

<sup>1</sup>Dunn, G., & Everitt, B. S., *An Introduction to mathematical taxonomy*, Cambridge University Press, 1982, p. 14 suggests a maximum of 100 characteristics and maybe 200-300 taxa. Collier, P. A., private communication, comments that the number of taxa can vary greatly from 2 to 800 or so, but

comparisons may be represented as a triangular matrix made up of the confidence levels with which the null hypothesis appropriate to the pair-wise comparison (that there is no difference between the two samples used in the comparison) may be rejected, e.g. in Figure 15,  $u\%$  is the level at which the null hypothesis (that there is no difference between samples  $\beta$  and  $\alpha$ ) may be rejected.

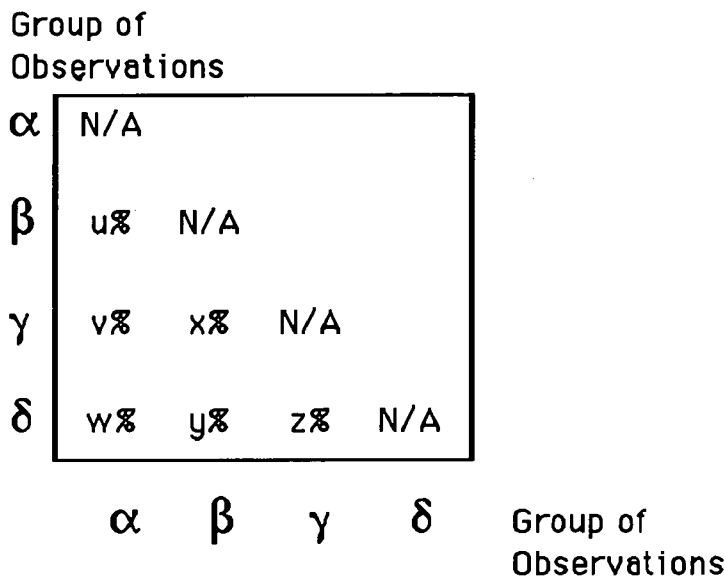


Figure 15 — Matrix of confidence levels.

A decision is then made about the reasonableness of possible splitting points, using the completed triangular matrix to check if the null hypothesis (that every group of observations below the proposed splitting point is the same distribution as every group of observations above this splitting point) can be rejected.

To give an example of this, the data represented in Figure 12 could be used to produce a triangular matrix of the type shown in Figure 16. For distribution  $\alpha$  of Figure 12 to be considered statistically distinguishable from distributions  $\beta$ ,  $\gamma$  and  $\delta$ , splitting point S2 would be used, and the results  $u\%$ ,  $v\%$  and  $w\%$  (see Figure 16) would all have to be less than or equal to 5% (or preferably, 1%).<sup>1</sup> If this were so, S2 could be used as a legitimate  $x_{split}$  value.

suggests the average would be less than 50 taxa. Pankhurst (1971) comments that 'keys to more than a few hundred taxa are rare ', (the type of reason being that illustrated in Table 4 of this thesis); ' even with a high probability of answering each lead correctly, the chance of a correct final result can be small, so that large keys are rather impractical' Pankhurst (1971).

<sup>1</sup>The maximum level that the program will report as an acceptable split is input as data at the start of a run. If the characteristics being used are expected to

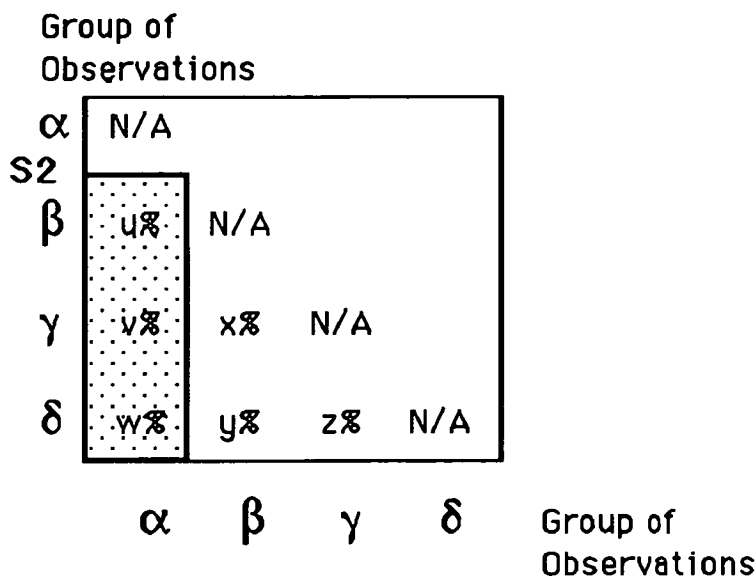


Figure 16 Splitting point S2 chosen.

If the splitting point S3 was chosen (Figure 17), the important results would be v%, w%, x% and y%. If all are below 1%, then this splitting point could also be regarded as reasonable.

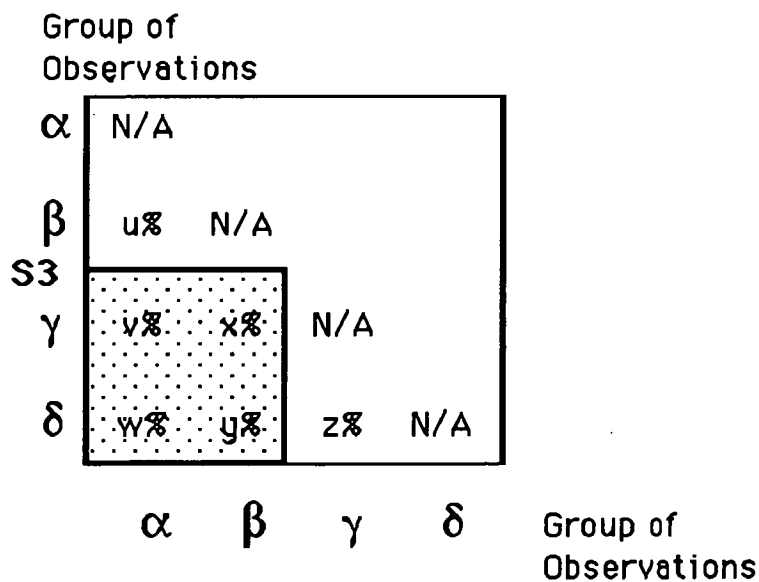


Figure 17 — S3 Splitting Point chosen.

If the splitting point S4 was chosen, (see Figure 18), the important results would be w%, y% and z%. If all are below 1%,

separate the species easily, a maximum level of 1% or lower could be used, and any level of split above this (e.g. 5%) would not be presented to the key developer. If the data is typical of botanic data, diffuse and not easily separated, a higher limit could be tried. However experience has agreed with theory that separations greater than 5% are rarely of use in practice.



(unlikely in this case, if Figure 12 is an accurate representation), then this splitting point could also be regarded as reasonable.

Group of Observations					
$\alpha$	N/A				
$\beta$	$u\%$	N/A			
$\gamma$	$v\%$	$x\%$	N/A		
S4 $\delta$	$w\%$	$y\%$	$z\%$	N/A	
	$\alpha$	$\beta$	$\gamma$	$\delta$	
	Group of Observations				

Figure 18 — Splitting point S4 chosen.

In cases where the level of rejection are different, the maximum of the appropriate rejection levels can be used as a 'figure of merit' characterising the splitting value chosen, a lower value indicating a stronger separation.<sup>1</sup> This could be used to

<sup>1</sup>An average of the appropriate rejection levels could also be used as a 'figure of merit'. However in the runs used in this thesis the maximum level of rejection  $m$  was defined by use of the fuzzy algebraic relationship (e.g. in the case of Fig. 16 and splitting point S2)  $m = u+v+w$ . (Kandel, Abraham and Byatt, William J., 'Fuzzy Sets, Fuzzy Algebra, and Fuzzy Statistics', *Proceedings of the I.E.E.E.*, Volume 66, No. 12, December 1978); . The choice of this type of split results in the worth of the split being judged by its weakest component, and is the most conservative of the several assumptions which could have been used. Fuzzy algebra was in used in this case with Zadeh's comment in mind, '...probability theory by itself or in combination with the maximum entropy principle, does not provide an adequate tool for analysis of problems in which the available information is incomplete, imprecise, or unreliable.' (Zadeh, Lofti A., 'Fuzzy Sets versus Probability', *Proceedings of the I.E.E.E.*, Volume 68, No. 3, March 1980). Smithson similarly argues for the use of fuzzy logic, noting that '...the most fundamental limitation of NPF [Neyman-Pearson-Fisher statistical framework] ... is the assumption that probability is capable of representing any form of uncertainty in human thought or behaviour.' (Smithson, Michael, 'Possibility Theory, Fuzzy Logic, and Psychological Explanation' in Zétényi, Tamás (Ed.), *Fuzzy Sets in Psychology*, North-Holland, Amsterdam, 1988, p. 5.) It is argued that uncertainty is, to some extent, inevitable. Chwedorowicz comments 'Every stimulus S reaching the subject can be presented as a pair (M, I) where M denotes message and I the information which it conveys. The same message, via various interpretations, can give various kinds of information (judgements). The difference in interpretation results from the different experience that different people have, and take the form of the so called "metabeliefs", beliefs about beliefs...' (Chwedorowicz, József, 'Origin, structure and function of fuzzy beliefs', in Zétényi, Tamás (Ed.), *Fuzzy Sets in Psychology*, North-Holland, Amsterdam, 1988, p. 269.) The effects of beliefs about beliefs (or axiomatic beliefs) are fundamental to the consideration of the ideas put forward in this thesis; examples of the effects of axiomatically held beliefs are discussed

choose between S2, S3 and S4. If required, several of the splitting points could be chosen simultaneously. In the case of Figure 12, it seems likely that both S2 and S3 are practical splitting points, resulting in a value of the variant being able to be used to separate distributions  $\alpha$  and  $\beta$  from the combined  $\gamma$  and  $\delta$ . To separate  $\gamma$  and  $\delta$  completely, another characteristic would have to be used.<sup>1</sup>

### 3.1.2.1.7 Type 1 errors and Decision Keys.

As discussed before, *1<sup>st</sup> Class* and other shells which use the general ID3 approach do not handle overlapping distributions at all well,<sup>2</sup> and this can lead to 'hidden' type 1 errors.

*1<sup>st</sup> Class* takes a similar approach in that it produces multiple exit points.<sup>3</sup> As an example, Figure 10 would contain some data of the type shown in Table 1. In the case of Table 1, *1<sup>st</sup> Class* would produce 8 separate conclusions rather than the two desired. Each group of scores corresponding to species A or B which either does not overlap or is equivalent will produce a separate conclusion; e.g. if extra readings (B, 2.35) and (A, 2.45) were included in Table 1, there would still be 8 conclusions as these readings would not increase the number of non-overlapping or equivalent groups; by contrast if readings (B, 2.5)

---

in section 1.5 of this thesis. These considerations also inevitably affect measurements taken as part of botanic data sets. As has been commented before, botanic data sets often contain specimens with incomplete measurements and measurements which are the result of an attempt to express an essentially qualitative characteristic in quantitative terms; (e.g. what proportion of green and blue, and what proportion of whitish bloom must be present for something to be classified as *glaucous*). In cases like this personal judgement and inclination inevitably are a factor in this type of classification, and the classification would probably be regarded as 'imprecise' by someone who had a background in the requirements for measurements in a discipline such as Physics. However it is essential when handling data of botanic origin that the methodology admits to the possibility of human judgement in the recording of the data. A similar caveat applies to data of both biological and psychological origin.

<sup>1</sup>Some methodologies exhibit a bias towards many-valued attributes, e.g. see: Quinlan, J. R., 'Induction of Decision Trees', in *Machine Learning*, Vol. 1, No. 1, p. 100. See also: Gower, J. C. and Payne, R. W., *A comparison of different criteria for selecting binary tests in diagnostic keys*, in *Biometrika*, Vol. 12 No. 3, 1965, p.671. The results obtained during the investigations reported in this thesis suggest that the methodology used in this thesis (*which included methodology appropriate to both parametric and non-parametric data*) does not seem to exhibit such a bias in the cases of the data tested, (although this conclusion has not been established *sensu stricto* and hence would profit from further investigation).

<sup>2</sup>Quinlan, J. Ross, *Simplifying Decision Trees*, Technical Report 87.4, New South Wales Institute of Technology, Sydney, 1987, pps. 3 - 16.

<sup>3</sup>Also referred to as "leaf nodes".

and (A, 2.6) were included the number of conclusions would increase to 10. In the case of Figure 10 which represents 400 test scores, the number of separate conclusions is likely to be very large, much greater than the number of actual conclusions (two). While the exact result can not be shown (as the data is summarised into classes in Garrett),<sup>1</sup> the number of conclusions is likely to number in the 10's rather than single digits, and may even exceed 100.<sup>2</sup>

Whilst the (ID3-type) methodology produces the problems shown above, the errors introduced by the cost-complexity, reduced error, pessimistic and other pruning methodologies discussed by Quinlan and others are real, but less obvious and often ignored in practice.

An expert, using *1<sup>st</sup> Class* to produce a botanical key, is tempted to reduce the number of conclusions by selectively omitting 'atypical' data. If the data is in fact not atypical, but is the result of an overlapping distribution with the overlapping portion under-represented by the sampling procedure used, the expert is in effect encouraging a type 1 error.

Consider the case shown in Figure 9.<sup>3</sup> Suppose an observation is measured as being above  $x_{split}$ . In this case it would probably be accepted that the object belonged to distribution  $\beta$ . If the object was in fact one of the rare objects belong to distribution  $\alpha$  which measured above  $x_{split}$ , a type 1 error would have occurred because the null hypothesis (that there was no significant difference between the observation and the values legitimately part of distribution  $\alpha$ ) would have been rejected when it was, in fact, true. By eliminating 'nuisance' or 'atypical' data which causes multiple exit points for a particular species, the expert is in effect reducing the effect of the portion of the distribution which results in type 1 errors. An extreme example of this would be to eliminate all  $\alpha$  data above  $x_{split}$ , and all  $\beta$  data below  $x_{split}$  in the case of Figure 9. This would make the data acceptable to *1<sup>st</sup> Class*, which could then produce a neat key; however the expert

---

<sup>1</sup>Garrett, pps. 73-74.

<sup>2</sup>See section 6.1.2, Figure 24 of this thesis for the outcome of a similar type of problem.

<sup>3</sup>See section 3.1.2.1.4 of this thesis.

has then 'built in' the type 1 error into the classification system in a way that is impossible for subsequent users to perceive or make any allowance for. In the case of the expert preparing the key, each breakpoint and separation decision is independent, the data being already classified. This is not the case when the key is employed by a user to identify a sample.

Considerations of this type limit the depth of the classification key that can be usefully employed in practice, depending on the size of the type 1 error that can be accepted. For example, to reliably achieve an expected classification rate in excess of 50% (using an acceptable null hypothesis rejection level of 0.8) the useful depth of a decision key is limited to about two questions, as shown in Table 4.

Null hypothesis rejection level	<i>Effect of type 1 error on effective depth of key</i>							
	(1 $\sigma$ ) 0.683	0.8	0.85	0.9	0.95	(2 $\sigma$ ) 0.955	(3 $\sigma$ ) 0.997	(4 $\sigma$ ) 0.999
Percentage correctly classified	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Depth 1	68	80	85	90	<b>95</b>	<b>95</b>	<b>100</b>	<b>100</b>
Depth 2	47	64	72	81	90	91	<b>99</b>	<b>100</b>
Depth 3	32	51	61	73	86	87	<b>99</b>	<b>100</b>
Depth 4	22	41	52	66	81	83	<b>99</b>	<b>100</b>
Depth 5	15	33	44	59	77	79	<b>99</b>	<b>100</b>
Depth 6	10	26	38	53	74	76	<b>98</b>	<b>100</b>
Depth 7	7	21	32	48	70	72	<b>98</b>	<b>100</b>
Depth 8	5	17	27	43	66	69	<b>98</b>	<b>100</b>
Depth 9	3	13	23	39	63	66	<b>98</b>	<b>100</b>
Depth 10	2	11	20	35	60	63	<b>97</b>	<b>100</b>

Table 4 — Percentage correctly classified after successive questions

Table 4 assumes that the 100 objects to be classified are of the same type, and each successive question has been chosen by a rejection of the null hypothesis at the level indicated, and that the questions are statistically independent. It will be noted that decisions taken below 2 $\sigma$  limits are risky, and lead to an (often unacknowledged) limit to the useful depth of the key. As an example, an expectation that 50% of the specimens should be correctly identified (allowing a level of null hypothesis rejection

of 0.9) would lead to a maximum useful depth of a key to 4-5 questions at best.<sup>1</sup> Since many botanical keys have depths within this range, it behoves any researcher deleting data points to examine carefully the statistical effect of the deletions. A possible criteria for rejecting an 'atypical' reading might be that it falls outside the  $\pm 3\sigma$  limits for the distribution which includes that reading.

If the resultant decision key was to be used in the production of rules for use in an expert system shell, it would be desirable to sum these errors and present the sum to the expert at the time the choice of question is made, as an indication of the reliability of the conclusion in the original data from which the rules were obtained. In the case of automatic key selection, the cumulative errors could be presented with the key. This way any type 1 errors would be "up front", and could be made visible to users of the key, instead of being invisible and leaving the key user uncertain whether the key was a strong one constructed without data deletions, or one which was approximate in that many deletions had to be made to produce a neat and useable result.

### 3.1.2.3 *Small Sample Parametric Tests.*

This methodology is appropriate when one can accept:-

- a) The assumption that the observations in the group are normally distributed<sup>2</sup>, and
- b) The requirement of a minimum sample size suggested for use with large sample normal distributions (30 examples<sup>3</sup> per observation group) is not met.

---

<sup>1</sup>There are parallels between this situation and Shannon's examination of transmission of information through a noisy channel. A high level of rejection of the null hypothesis (e.g. > 5%) could be said to correspond to what Shannon calls a high level of equivocation or conditional entropy of the received signal, (see Shannon, Claude E. and Weaver, Warren, *The Mathematical Theory of Communication*, 12<sup>th</sup> Edition, University of Illinois Press, Urbana, U.S.A., September 1949, p. 67). Too deep a key constructed at too high a level of rejection of the null hypothesis could lead to a situation where the level of identification of a species using such a key might not exceed that achievable by chance.

<sup>2</sup>Garrett, p. 105.

<sup>3</sup> Garrett, p. 215; Kohler, Heinz, *Statistics for Business and Economics*, Scott, Foresman and Company, London, 1985, p. 300, 312 concurs, but adds that size of sample should be  $< 0.05(\text{size of population})$ .

The following discussion of the small sample methodology assumes that the previous section 3.1.2.1 has been read. The discussion is therefore much briefer than that in section 3.1.2.1. Section 3.1.2.3.1 discusses two distributions which may be appropriate in this case, and section 3.1.2.3.2 presents the mathematical background necessary to enable a splitting point to be chosen. The next two sections (3.1.2.3.3 & 3.1.2.3.4) note the modifications necessary to the large sample approach to fit the small sample problem.

### 3.1.2.3.1 *Introduction*

If the number of observations in a group or set is less than about 30, even if the observations are drawn from a normal or Gaussian population, special precautions must be taken. Two distributions are widely used in this case, the  $t$  and  $\chi^2$  distributions. We will mainly use the  $t$  distribution, preferring it as more appropriate than the  $\chi^2$  test for small samples (less than 30)<sup>1</sup> which are normally distributed. The  $t$  distribution is similar in form to the normal curve shown in Figure 4, with the exceptions that the maximum is lower and the 'wings' on either side are higher, (i.e. the standard deviation is greater). As the number of observations increases the maximum increases and the 'wings' lower, with the distribution being virtually the same as the normal distribution when the number of observations is equal to infinity.<sup>2</sup>

The following discussion of small sample parametric distributions is briefer than the preceding discussion, as the concepts are similar to those in the foregoing discussion of large sample distributions.

### 3.1.2.3.2 *Difference between means, small sample parametric distributions*

Let us consider the situation shown in Figure 7,<sup>3</sup> postulating two distributions where  $n_\alpha$  is of the order of 6,  $n_\beta$  is about 10. If it is desirable to see if the means of the two distributions are

---

<sup>1</sup> $\chi^2$  tests prefer >30, preferably > 100 examples, see Gryna, Frank M., *Basic Statistical Methods* in Juran, Gryna & Bingham, pps. 22-44.

<sup>2</sup>Garrett, p. 192

<sup>3</sup>Figure 7 is in section 3.1.2.1.4 of this thesis.

sufficiently separated from each other for the null hypothesis (that both are drawn from the same distribution) to be rejected, the following formulæ apply.

Firstly a joint standard deviation is calculated, which will apply to each of the groups, see equation 14.<sup>1</sup>

$$\sigma_{\alpha\beta} = \sqrt{\frac{\sum_{i=1}^{n_{\alpha}} (x_i - \mu_{\alpha})^2 + \sum_{j=1}^{n_{\beta}} (x_j - \mu_{\beta})^2}{(n_{\alpha} - 1) + (n_{\beta} - 1)}} \quad (14)$$

The standard deviation of the mean,  $\sigma$ , is found by equation 15.

$$\sigma = \sigma_{\alpha\beta} \sqrt{\frac{n_{\alpha} + n_{\beta}}{n_{\alpha} n_{\beta}}} \quad (15)$$

Now a critical ratio, (t), is found from equation 16.

$$CR = \frac{\mu_{\beta} - \mu_{\alpha}}{\sigma} \quad (16)$$

Most statistical texts will give the critical ratios for 5% and 1% levels of significance corresponding to the number of degrees of freedom (df) appropriate to distributions  $\alpha$  and  $\beta$ ,<sup>2</sup> where:-

$$df = (n_{\alpha} - 1) + (n_{\beta} - 1) \quad (17)$$

If the CR is greater than the 5% or 1% figure obtained from the text, then the null hypothesis that both distributions  $\alpha$  and  $\beta$  are drawn from the same population can be rejected.

---

<sup>1</sup>Garrett, p. 224; Note that this version of the t test implicitly assumes the two distributions have similar standard deviations. If the standard deviations are different, (see Kreyszig p. 965 for an appropriate test), a non-parametric test could be used, see section 3.1.3 of this thesis. Alternatively, the t distribution could still be used in a form which allows for variation between the respective standard deviations, e.g. see Gryna, Frank M., *Basic Statistical Methods* in Juran, Gryna & Bingham, p. 22-439 or Steele & Torrie, p. 106 for an appropriate form.

<sup>2</sup>E.g. Popham, W. James, and Strottnik, Kenneth A., *Educational Statistics, Use and Interpretation*, 2nd Edition, Harper & Row, New York, 1973, p. 38; or Kohler, p. T-28; or Kreyszig, p. A-69.

### 3.1.2.3.3 *Choosing a splitting point, with small sample distributions.*

If the approach above is taken, and a *joint* standard distribution  $\sigma_{\alpha\beta}$  is calculated which refers to *each* of the two groups, all of the alternatives for choosing splitting groups examined in the case of large sample parametric distributions reduce to the situation represented by equation (9).<sup>1</sup>

While separate  $\sigma_{\alpha}$  and  $\sigma_{\beta}$  figures can be obtained by using the large sample methods already discussed, their statistical validity is somewhat doubtful, (particularly if  $n_{\alpha}$  and  $n_{\beta}$  are less than 10). Thus while it probably would not hurt to use large sample techniques for the choice of  $x_{split}$ , it is probably quite adequate to simply set  $x_{split}$  to the mid-point between the means  $\mu_{\alpha}$  and  $\mu_{\beta}$ .

### 3.1.2.3.4 *Distinguishing between many small-sample distributions*

This is the same in principle as the method described in section 3.1.2.1.6 for large-sample distributions, and so will not be repeated here.

## 3.1.3 Non-Parametric Tests

The small and large sample parametric tests described above are strictly only applicable if the user does not reject the null hypothesis that there is no significant difference between the distribution of the data being used and a normal distribution. If the null hypothesis can be rejected, then non-parametric tests are applicable. Section 3.1.3.1 notes several non-parametric tests which may be appropriate. Section 3.1.3.2 examines a way in which one of these options, the randomisation test, may be of use. Section 3.1.3.3 provides a very brief summary of the applicability of non-parametric tests to key generation.

---

<sup>1</sup>See section 3.1.2.1.4 of this thesis.



### 3.1.3.1 Introduction to non-parametric tests.

If the data to be examined fails a test for normality,<sup>1</sup> there are several non-parametric or distribution-free tests which could be considered. The following sections consider the Sign test which 'is the simplest and most generally applicable of the non-parametric tests'<sup>2</sup> (section 3.1.3.1.1) and the U test which 'has been found to be very useful for testing hypotheses'<sup>3</sup> (section 3.1.3.1.2); both tests being examined because they are amongst the first usually considered by psychologists when faced with the task of analysing non-parametric data. Section 3.1.3.1.3 introduces the idea of randomisation tests, and 3.1.3.1.4 considers some of the advantages and disadvantages of this latter type of test.

#### 3.1.3.1.1 Sign Test

The Sign test is useful in that it makes no assumption about the shape of the distribution.<sup>4</sup> However it does assume data is presented as paired sample values. Hence with botanic data such as the *Acaena* data, where sample sizes vary from species to species, some of the data would have to be discarded in each comparison. Also the test uses only the sign of the comparison of values, and information about the magnitude of the differences is not used. These last two factors make the test the least powerful of those considered for use in a distribution-free version of "Selecta-key".

#### 3.1.3.1.2 U Test

The U test employs the rank order of the data, and so (unlike the sign test) takes some account of the magnitude of the statistics.<sup>5</sup> This makes it potentially more powerful than the sign test. However, while the distribution of the variables is not assumed to be normal, the U test does make an assumption that both the distributions are from the same population — i.e. that

---

<sup>1</sup>Such as the Kolmogorov-Smirnov test for normality, see Gryne in Juran, Gryna & Bingham, pps. 22-44; this test is also implemented in the computer program Statworks which runs on the Macintosh computer.

<sup>2</sup>Garrett, p. 267.

<sup>3</sup>Hoel, p. 291.

<sup>4</sup>Hoel, p. 285.

<sup>5</sup>Hoel p. 291. See also Melsa, James L. and Cohn, David L., *Decision and Estimation Theory*, McGraw-Hill, New York, 1978, p. 169 for details of the Wilcoxon rank order test.

the shape of the distribution (whatever it was) of each of the two sets of data was the same. This seems to be an unreasonably restrictive assumption, given that the assumption of the applicability of the most commonly occurring (normal) distribution has not been able to be upheld. A test which makes no assumptions at all about distribution shape would be preferable.

### 3.1.3.1.3 *Randomisation Tests — Introduction*

Randomisation tests have the advantage of making no assumption whatsoever about the shapes of the distributions of the data being examined. However they do have one marked disadvantage when compared with the more generally used parametric experiential methods; the results obtained are, *sensu stricto*, not applicable beyond the data examined.<sup>1</sup>

In the physical sciences, experimentation has traditionally meant careful direct control over extraneous variables, and both precise manipulation of variables and precise measurements of the effects of those manipulations. The subjects or objects upon which the experimentation occurs have often been carefully chosen by random selection from a larger population, the careful choice enabling any conclusion which applies to the experimental sample to also apply (within probabilistic limits) to the population as a whole.

In contrast, the randomisation tests employ random *allocation*, rather than the random *selection* employed by more widely used tests. The test statistics are repeatedly randomly allocated, and conclusions can be drawn from these allocations.

### 3.1.3.1.4 *Randomisation Tests — Advantages & Disadvantages*

As noted in section 2.2.3 of this thesis, it is not unusual that data of botanical origin does not meet the statistical standards for random sampling of a population. Randomisation tests have

---

<sup>1</sup>This is the same limitation which would occur if a statistically unrepresentative sample of a parametric distribution was used. In practice this could be a marked limitation in work where it is possible to obtain a carefully planned statistically representative sample of the larger population under investigation (as is possible in many psychological investigations), but is less of a limitation in botanic work where it is more common that distance, geographical distribution and expense can make gathering such a statistically valid representative sample very difficult to impossible in practice.

the advantage that they can legitimately draw conclusions from this sort of data. Edgington comments:

*Statistical inferences cannot be made concerning populations that have not been randomly sampled; therefore, few experiments would be published if it were necessary to show that the experiment permitted a statistical inference concerning an important population, a population of general interest to the readers. We will now argue that random sampling of a population is not relevant to most psychological experimentation and that the lack of a random sample does not prevent drawing useful statistical inferences - about the experimental subjects actually used.*<sup>1</sup>

Thus if the expert feels strongly that data which does not meet the statistical standard for random sampling is representative of the population as a whole, he or she may make a case for generalising the results — however it must be stressed that this generalisation is a *non-statistical generalisation*, not based on the statistical method, but on the expert's knowledge of how typical is the data that has been fed into the expert or key construction system. (The same restriction would appear to apply to some other non-statistical systems using this type of data, e.g. neural nets, ID3, clustering, to name a few).

Restrictions of this type may apply in practice much more widely than the statistical methods employed by researchers would suggest. This disadvantage was felt to be significant in theory, but not serious enough in practice to disqualify this test from consideration.<sup>2</sup>

The second disadvantage is that random allocation, the so-called monte carlo methods, are sometimes regarded by some as brute force methods lacking in theoretical validity. In view of

---

<sup>1</sup>Edgington, pps. 96-97. The italics were in the original text.

<sup>2</sup>In the case of the type of collections of botanic specimens to be considered later in this thesis, there can be significant limitations in the researcher's ability to collect a representative sample. Practical considerations, such as the number of days of food that can be carried in the collector's backpack, or the paucity of research funds to enable the employment of assistants to carry out detailed measurements not within the abilities of the collector (either because of time or skill restrictions, or the unavailability of specialised equipment), combine to make collection of a representative sample of data generally less than ideal in practice. These type of problems are considered in greater detail in section 2.2.3 of this thesis.

this, the author was interested to note Forsyth's call for increased use of randomisation methods. It is felt that in this case their use is justifiable.<sup>1</sup>

The third disadvantage is that repeated random allocation takes time.<sup>2</sup> This has been a powerful objection in the past, however with the advent of high power computers which cost considerably less and which are more widely available than they have been in the past this is felt to be much less of a disadvantage than previously.<sup>3</sup>

This method could be expected to be much slower in practice than the application of either the normal or student's t tests used in the large and small sample tests already discussed. This could be a marked disadvantage if this test were to be used every time.

### 3.1.3.2 Randomisation Tests — Possible method of use.

Edgington comments that 'a normal curve test can sometimes be used as an approximation to the randomisation test'.<sup>4</sup> This comment, probably based on the central limit theorem,<sup>5</sup> suggests that the way the total program could be used would be to attempt to obtain a key assuming that either the normal or student's t test was applicable. If the result looked potentially useful, the data could then be checked to see if the null hypothesis (that the data was not significantly different from a normal distribution) could be rejected. If the null hypothesis was not rejected, the classifications and keys produced by the program could then be accepted. If some of the groups of data failed the null hypothesis, then a randomisation tests could be run to check the results.

---

<sup>1</sup>Forsyth, R. S., 'The Evolution of Intelligence', in Third International Expert Systems Conference, Learned Information Ltd. (Ed.), London, 1987, pps. 61 - 75.

<sup>2</sup>This objection has also been levelled at some other methodologies, e.g. neural nets.

<sup>3</sup>E.g. Levco's announcement of a 200 million instructions per second upgrade for the Macintosh 11, see MACazine, DATELINE: Macintosh, Icon Concepts Corporation, Austin Texas, October 1987, p. 111; and the 2.5 milliard instructions per second connection machine, see Hillis, W. Daniel, 'The Connection Machine', *Scientific American*, Scientific American Incorporated, New York, June 1987, Vol. 256, No. 6, pps. 86-93.

<sup>4</sup>Edgington, Eugene S., *The Distribution-free approach*, McGraw-Hill, New York, 1969 p.161.

<sup>5</sup>op.cit. p 73

Adoption of this methodology means that randomisation tests could be fitted into the Selecta-key methodology, and they were chosen as the preferred non-parametric test. Section 3.1.3.2.1 uses an example to discuss the randomisation test methodology. The groups have a minimum size, (section 3.1.3.2.2), and whilst theoretically there is no maximum size, in practice a combinatorial explosion limits the size of the groups, unless the field of possible randomisation arrangements is sampled (section 3.1.3.2.3). Section 3.1.3.2.4 compares the randomisation test with parametric tests. Section 3.1.3.2.5 examines the proposition that randomisation tests performed on a computer using the original data would be better than a key. Section 3.1.3.2.6 comments on the approach adopted by Selecta-key. Sections 3.1.3.2.7 and 3.1.3.2.8 note a suitable method for selecting a splitting point for single and multiple characteristics, respectively.

#### *3.1.3.2.1 Distinguishing between two non-parametric distributions*

Let us now examine an example of the application of the randomisation test. The test is probably best explained by use of an example. Consider the problem of checking if the null hypothesis (that the *Acaena echinata* var. *robusta* and *Acaena agnipila* var. *protenta* leaf length data is from the same distribution) can be rejected.

The values observed by Collier are shown in Table 5, following:-

LEAF		LENGTH	
<i>Acaena echinata</i> var. <i>robusta</i>		<i>Acaena agnipila</i> var. <i>protenta</i>	
		17.5	
8		8.5	
11		15.0	
Average = 9.5		Average = 13.7	

Table 5 — Observed values of leaf length.

The object of the test is to find out if there is any significant difference between the leaf lengths recorded for the two species. This could be tested by finding out if the measurements for one species are the same as they would have been had they belonged to the other species. In this case we have in total five measurements — not drawn at random from the entire population, but the only measurements currently available, (and in the case of *Acaena agnipila* var. *robusta*, possibly the only measurements likely to become available, as this species has been observed very rarely, is probably extinct, and the two samples measured are the only examples of this form known to the author. Even if more do become available, the point remains valid that there are many botanical species rare enough to give researchers similar problems.)<sup>1</sup>

If the supposedly two populations were, in fact, drawn from the one population, then any difference in average between the sample of 2 (*Acaena echinata* var. *robusta*) and the sample of 3 (*Acaena agnipila* var. *protenta*) would be solely due to the type of differences which could be caused by a random allocation of the five measurements to the two groups. The number of different

<sup>1</sup>This is a situation which is familiar to statistical practitioners.; e.g. Macnaughton-Smith comments: 'The reader will be familiar with cases where even a totally enumerated finite population (such as total admissions to a given institution in a given period) is regarded for research purposes as a random sample from the underlying infinite population of all the cases which could have arisen in the given situation'; Macnaughton-Smith, P., *Some statistical and other numerical techniques for classifying individuals*, Her Majesty's Stationery Office, London, 1965, p. 2.

ways of randomly allocating (n+m) measurements to two groups of size n and m is equal to:-

$$= \frac{\overline{n+m}}{\overline{n} \overline{m}} \tag{18}$$

In this case, number of possible ways of randomly allocating the five measurements to the *Acaena agnipila* var. *protenta* and *Acaena echinata* var. *robusta* groups is equal to ten. All ten allocations are listed in Table 6, where the first row lists the observed measurements. The other nine rows of this Table use the same five measurements randomly allocated to the two species groupings.

<i>Acaena echinata</i> var. <i>robusta</i>		<i>Acaena agnipila</i> var. <i>protenta</i>			<i>Acaena agnipila</i> var. <i>protenta</i>
Leaf	Lengths	Leaf Lengths			Sum of Leaf Lengths
8.0	11.0	17.5	8.5	15.0	41.0
17.5	8.0	8.5	15.0	11.0	34.5
8.5	17.5	15.0	11.0	8.0	34.0
15.0	8.5	11.0	8.0	17.5	36.5
11.0	15.0	8.0	17.5	8.5	34.0
8.0	8.5	17.5	15.0	11.0	43.5
8.0	15.0	8.5	17.5	11.0	37.0
8.5	11.0	15.0	8.0	17.5	40.5
17.5	11.0	15.0	8.5	8.0	31.5
17.5	15.0	11.0	8.0	8.5	27.5

Table 6 — Randomisation of two *Acaena ovina* measurement groups

Some statistic associated with each allocation must now be defined. Possible candidates include the average leaf length, the length squared, the log of the length, plus several other statistics which could have been chosen. However all would have produced

the same result, as the rank order for all would have been the same. In this case the sum of the lengths allocated to *Acaena agnipila* var. *protenta* was chosen as the testing statistic, simply because it was easy to calculate. This sum is listed in the last column of Table 6.

The prediction to be tested would be that the observed *Acaena agnipila* var. *protenta* sum would be greater than the sums produced by random allocation of the measurements to the two groups. The corresponding one-tailed null hypothesis would be that for no leaf would there be an *Acaena agnipila* var. *protenta* measurement larger than the *Acaena echinata* var. *robusta* measurement. When the measurements are randomly assigned into equally probable sets of data, the test actually corresponds to a two-tailed test, however Edgington<sup>1</sup> proves that the probability under the one-tailed null hypothesis will be no higher than the probability under the two-tailed null hypothesis that the size of the leaf would have been the same regardless of the "species" group from which it had been drawn.

If the null hypothesis that the two groups are drawn from the one collection could not be rejected, each of these random allocations would be equally likely, and any difference between the groups would be the results of 'randomisation error', discrepancies resulting from the random assignment of the measurements.

In this case, inspection of Table 6 shows that in one case the 'sum of each *Acaena agnipila* var. *protenta* measurements' (allocation 5) is greater than the sum of the observed measurements, i.e. the number of allocations greater than or equal to the observed measurements is 2. Thus in this case the probability under the two-tailed null hypothesis of getting a set of results equal to or greater than 41 is 2/10; under the one-tailed null hypothesis the probability is no greater than 2/10. Thus the result of the grouping would not be significant at the 0.05 level, but it would be at the 0.20. Hence this data would be inadequate to reject the null hypothesis. The data presented could not reasonably be used to indicate that the leaf length of *Acaena*

---

<sup>1</sup>Edgington pps. 137 - 138.



*agnipila* var. *protenta* is significantly greater than the leaf length of *Acaena echinata* var. *robusta*.

### 3.1.3.2.2 Randomisation Tests — Minimum group sizes

It will be noted that this result could have been predicted before the test was run, as 20 or more random allocations are required before rejection at the 0.05 level can be achieved. The minimum number of allocations for small class sizes is given in Table 7.

Total members	Number of Members		Number of allocations
	Group A (n)	Group B (m)	
20	1	19	20
7	2	5	21
6	3	3	20
7	4	3	35
7	5	2	21
8	6	2	28
.	.	.	.
.	.	.	.
20	18	2	190
20	19	1	20

Table 7 — Minimum Group Size for Randomisation Test.

Usually the problem is not that the number of random allocations is too small, (as in the previous example), rather the problem is that the number of allocations is too large to be computationally convenient using a desk-top computer.

### 3.1.3.2.3 Approximate Randomisation Tests

Large groups create a problem for randomisation tests, if complete examination of all alternatives is required. For example, in the last example in Table 7, if group B had the same number of members as group A (19), the number of random allocations possible would not be 20, but approximately  $3.5 \times 10^{10}$ . This is an inconveniently large figure. Some method of sampling this huge range of possibilities is needed. The sample chosen must be small enough to be calculable in a reasonable time, but large enough to achieve a reasonable degree of probabilistic

prediction that the result attained applies to the whole set of measurements.

In Section 3.1.3.2.3.1 work by Edgington is used to outline a sampling methodology which could be used for dealing with very large groups. Section 3.1.3.2.3.2 presents a similar method for use with smaller groups. In both cases the resulting sample size is manageable, and can be treated by the same process outlined in section 3.1.3.2.1. It is also shown in these sections that there is a reasonable probability that the sample size recommended reflects the behaviour of the distribution from which it is drawn.

#### 3.1.3.2.3.1 *Approximate Randomisation tests for large groups*

Edgington suggests a methodology for dealing with large groups, and refers to the process of randomly selecting samples from the entire sampling distribution and the resulting test as 'approximate randomisation tests'.<sup>1</sup> He proposes drawing 999 samples randomly from the entire sampling distribution to add to the one obtained statistic, obtains the fiduciary limits for the significance level, and makes the following statements:-

The probability is .99 that an obtained statistic that would be judged significant at the 0.01 level by using the entire sampling distribution will be given a probability no greater than .018 by using the approximate sampling distribution.<sup>2</sup>

and

The probability is .99 that an obtained statistic that would be judged significant at the .05 level by using the entire sampling distribution will be given a probability no greater than .066 by using the approximate sampling distribution.<sup>3</sup>

Edgington derived these statements in relation to correlation statistics, but comments that they apply to any randomisation test statistic. He obtains these results by applying the binomial<sup>4</sup>

---

<sup>1</sup>Edgington, pps. 152 - 155.

<sup>2</sup> op.cit. p.154.

<sup>3</sup> op.cit. p.155.

<sup>4</sup>Garrett, Henry E., and Woodworth, R.S., *Statistics in Psychology and Education*, Vakils, Feffer and Simons Pty. Ltd., Bombay, 1967. pps. 89-94.

or Bernoulli<sup>1</sup> distribution. The properties of this distribution can be mathematically represented<sup>2</sup> as in equations 19, 20 and 21:-

$$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad (19)$$

where:-

$p(x)$  = the probability of exactly  $x$  successes in  $n$  trials  
 $p$  = the probability of success in one trial.

$$\mu = np \quad (20)$$

$$\sigma = \sqrt{np(1-p)} \quad (21)$$

where:-

$\mu$  = the distribution mean  
 $\sigma$  = the distribution's standard deviation

It is also possible to derive an approximate formula for the minimum number of samples  $n_{min}$  needed for an assurance that the probability of rejection is in a certain range  $\delta$ .<sup>3</sup>

$$n_{min} = \frac{Z^2 P(1-P)}{\delta^2} \quad (22)$$

where:-

$z$  = the number of standard deviations from the mean

e.g. For convenience in calculation,  $z$  is usually taken as 2, i.e. to obtain a 95.4% ( $z=2$ ) confidence level that the probability of rejection was to be  $0.05 \pm 0.02$ ,  $n_{min}$  would be 475.

A complication with this approach is that the binomial distribution assumes that the size of the overall distribution from which the samples are taken is infinite. (An alternative formulation for the Bernoulli distribution, that the selections are made *with replacement*, does not apply in this case as it is not desirable that a sample random allocation be used more than once). Obviously this requirement of an infinite sample space is not met in practice, but Burr comments that if the overall

<sup>1</sup> Hoel, pps. 62 - 64.

<sup>2</sup> Ali, A.M., 'Probability - Uncertainty - Simulation', in Jelen., F.C., *Cost and Optimisation Engineering*, McGraw Hill Book Company, New York, 1970, pps. 156 - 157.

<sup>3</sup> Obtained by a combination of Edgington, p. 161 and Kohler, p. 335.

distribution is 8 to 10 times the size of the sample, then a reasonable approximation is obtained.<sup>1</sup> Other authorities suggest the overall distribution should at least be 10 times<sup>2</sup> to 20 times<sup>3</sup> the sample size.

### 3.1.3.2.3.2 Approximate Randomisation tests for small groups

If the overall distribution is too small, Jelen suggests 'the binomial distribution should be replaced by the hypergeometric distribution'.<sup>4</sup> In this case the formulæ for the mean  $\mu$  and standard deviation  $\sigma$  are those shown in equations 23 and 24.<sup>5</sup>

$$\mu = n * \left( \frac{M}{N} \right) \quad (23)$$

$$\sigma = \sqrt{np(1-p) \left( \frac{N-n}{N-1} \right)} \quad (24)$$

where:-

- N = number of objects in the entire collection
- M = number of the N objects which are successes
- n = number of objects in the sample
- p = probability that an object in whole collection is a success, ( = M/N).

### 3.1.3.2.4 Randomisation Tests — Comparison with parametric tests

In summary, it will be noted that the randomisation test on the two species of *Acaena* is quite different from the large and small sample parametric tests. In both these latter cases the parameters (means and standard deviations) were established first, and a test for distinguishability made in terms of these parameters. In the case of the randomisation test, the test distinguishability was made without any initial calculation of distribution parameters.

<sup>1</sup>Burr, Irving W., *Engineering Statistics and Quality Control*, McGraw Hill, New York, 1953, p. 201.

<sup>2</sup>Gryna, in Juran, Gryna and Bingham, pps. 22-19.

<sup>3</sup>McPherson, D.G., University of Tasmania, untitled and undated lecture handout, page 9.8.

<sup>4</sup>Jelen p. 157; also Kreyszig, p. 924; background theory can be found in Carlson, B. C., *Special Functions of Applied Mathematics*, Academic Press, New York, 1977, p. 14.

<sup>5</sup>Burr, p. 203; also Kreyszig, p. 925.

As in the case of the normal distributions, pair-wise comparisons would need to be used. (Tests are available for multi-modal distributions, but they require very large data samples to be effective in practice. This size of sample is rarely available to a botanist intent on constructing a key, so this avenue was not pursued).

#### *3.1.3.2.5 Randomisation Tests — Are keys needed?*

For a 'pure' application of non-parametric tests to species identification by an expert system, the data would be checked for distinguishability, and if acceptable, the data would be dumped into a data base for use by the expert system, which could then simply request data for each characteristic of a specimen and inductively identify that specimen from the data base by using randomisation tests rather than using any splitting points. This would have the disadvantage of being computationally very intensive, but would make the limitations inherent in the data base (normally hidden in rules derived from the data base) apparent to the user.

However the use of splitting points and the possible production of a key, does provide some advantages in practice, the advantages being:-

- the resultant key would be available to botanists in paper as well as computer-based expert system form;
- an identification of species by the expert system would be much faster;
- the expert system would be available on a smaller (and hence usually cheaper) computer;
- the basis for identification would be clearer, and more accessible to the human expert.

The main disadvantage of working from a key (rather than from the original data) is that identification of an unknown species by the resultant expert system would be theoretically less accurate, as in 'pure' randomisation the expert system would work from the original data every time an identification was required.

### 3.1.3.2.6 *Randomisation Tests — Approach adopted*

It was decided to implement Edgington's approximate randomisation test as part of Selecta-key, to handle cases which it would be considered unwise to classify as statistically normal.

### 3.1.3.2.7 *Randomisation Tests — Splitting Point Selection.*

Large sample parametric distribution methods use mean and standard deviation information to estimate the splitting point between distributions. In this case a corresponding measure of distribution spread could be obtained either by the same method and equations 9 or 10, or by considering either a cumulative frequency graph or an ogive, and finding the 2.5<sup>th</sup>, 50<sup>th</sup> and 97.5<sup>th</sup> percentile points. In this case the 50<sup>th</sup> percentile would give the median (the measure of central tendency corresponding to the normal distribution's mean) and it would be known that 95% of the distribution would lie between the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentile points. It will be noted that the difference between the 97.5<sup>th</sup> and 50<sup>th</sup> percentile in terms of the characteristics being measured may not be the same as the difference between the 2.5<sup>th</sup> and 50<sup>th</sup> percentile. Hence the measure of spread of the non-parametric distribution may not be symmetrical about the measure of central tendency (as is the case with the normal and t distributions). If this is taken into account, similar methods to those suggested for the large-sample parametric distributions could be implemented using the percentile measures of spread and central tendency. This may be useful if the number of observations in the distribution is large, and the distribution very skewed.

If the number of observations is small, then probably the approach taken by equation 10 (preferred) or 9 is adequate.<sup>1</sup>

### 3.1.3.2.8 *Randomisation Tests — Distinguishing between many distributions.*

This problem can be handled in a way which is similar in principle to the approach taken for large-sample parametric distributions, with the exception that the tests are made using the randomisation methods discussed herein, with the number

---

<sup>1</sup>See equations 9 and 10 in section 3.1.2.1.4.

to be chosen in each group being set by the number above the splitting point in the total group formed by combining the pair under consideration.

### 3.1.3.3 Randomisation Tests — Summary

The general principles used with large-sample and small sample distributions can also be applied to non-parametric distributions, albeit at the cost of significantly more computation.

All of these methods seem practical for use in a Selecta-key system.

## 3.2 Error Correction — Use of Multiple Characteristics.

The previous discussion of parametric and non-parametric tests assumes the use of only one characteristic. In general the preferred practice would be to use multiple characteristics per key decision. Thus each decision in the key would either be made with the a single question involving the "best" single characteristic (section 3.2.1) or (preferably) the "best" single question plus multiple other questions involving the use of other characteristics. The key could then be constructed (section 3.2.2). The additional use of multiple characteristics could provide a measure of error correction (depending on the number of questions asked regarding that key decision, see section 3.2.3).

### 3.2.1 Choice of the best single characteristic to use

After examining all the characteristics available, one would be chosen. If selection was performed by an expert, in association with information supplied by Selecta-key, there would hopefully be a much higher chance of an understandable and useful decision key resulting than if the decision was made automatically. The information to be supplied by a Selecta-key system could include:-

- a) the total number of splitting point options available for the expert to choose from, and the serial number of the option at present displayed on the screen;

- b) the characteristic used by the splitting point  $x_{split}$ ;
- c) the value(s) chosen by Selecta-key for  $x_{split}$ ;
- d) the taxa separated, shown in some sort of diagrammatic form, so the information would be immediately available to experts with inductive logic as well as deductive logic.
- e) Some quantification of how "good" the currently suggested splitting point is, (to permit comparison with the other splitting points currently on offer).

Since there would generally be many candidates for  $x_{split}$ , these could be ranked by Selecta-key, (e.g. the ranking could be in order of strongest split chosen from the split(s) (if any) available in the characteristics being considered). These splitting points could then be presented in turn, the rate of presentation preferably being controlled by the expert.

The expert could then choose the preferred characteristic and corresponding  $x_{split}$ , and in this way make allowance for peculiarities of the data known to the expert, such as characteristics which may either be seasonal (e.g. flowers or fruit data) or difficult to measure (e.g. items which require use of specialised equipment to obtain the necessary measures).

The choice of characteristics may also depend on the purpose for which the key is intended; e.g. different keys may be necessary for particular times of the year, (e.g. leaf dimensions may not be available if deciduous species are being examined during winter).

An expert using Selecta-key could hopefully meet these needs without having to edit or possibly re-specify participating characteristics between runs, as is required by *1<sup>st</sup> Class*.<sup>1</sup>

---

<sup>1</sup>Since *1<sup>st</sup> Class* generated a decision tree automatically, the only way to get a different decision tree was to either edit the data to omit some of the data, or re-arrange the data in a different order, and hope the new key would be what the user required. By contrast, when producing an alternative key using Selecta-key, the data was not changed; the Expert simply selected from the splitting choices presented by the program to obtain the appropriate decision tree. This latter approach, as well as being less effort, not only eliminated any possibility of data error creeping into the data due to erroneous editing, it also required less computer expertise in the user (important when the program is being used by an expert from another discipline).



### 3.2.2 Construction of the key

The key could be constructed in either of two ways; the Expert choosing options aided by Selecta-key (section 3.2.2.1), or automatically (section 3.2.2.2).

#### 3.2.2.1 Key construction — *Expert aided by Selecta-key*

After having chosen the first split by either of the methods mentioned above, the Expert could use Selecta-key to calculate the subsequent splits in a similar manner. The second question the system asks the Expert would be chosen assuming the alternative chosen for the first, and re-counting the number of statistically significant pair-wise comparisons that are now appropriate. Note that these would not have to be recalculated in most cases, merely the appropriate ones chosen and re-counted. (The figures-of-merit which are necessary to rank the alternatives would, however, have to be recalculated.) Other characteristics which may be appropriate as additional splitting criteria could be displayed by the program for selection or rejection by the expert (these can be identified easily, as they have the same taxa split as that chosen for the primary split).

This approach would continue with the third and subsequent questions, resulting finally in a decision key. The resultant decision key should be able to be printed out in an acceptable format.

#### 3.2.2.2 Key construction — *Automatic*

Automatic key construction is not of interest to this thesis, for reasons outlined more fully in section 2.2.3. However since this approach is unusual in the area of A.I., it may be germane to briefly re-state the observations of an authority who is regarded as one of, if not the, world authority in the area of taxonomic key construction by computer:

Batch mode key-construction programs have been in use for as long as twenty years, but have not found universal acceptance. Evidence has accumulated that keys produced by batch methods are still regarded as being less than ideal. ... This would be true for any computer-constructed key. ... Taxonomic

experts prefer to make subjective choices of characters at every stage ... The discussion attached to the review of Payne and Preece (1980) shows that taxonomists, mathematicians and computer programmers differ on this point.<sup>1</sup>

Pankhurst then makes an important point which is vital to the approach taken in this thesis:

The purpose of an interactive key-constructing program is therefore not to increase mathematical refinement in the algorithms but to increase the participation by the taxonomic expert.<sup>2</sup>

Although it is thus irrelevant to the purpose of this thesis, it may be of academic interest that, given a trivial alteration to the program, keys could have been generated entirely automatically, with selection of splitting points being made on the basis of the selection of the first ranked option which would have been presented to the expert, producing a monothetic key which would have been statistically valid to a level selected by the expert before the run. A less trivial, but still simple alteration, would have allowed the automatic selection (where appropriate) of multiple characteristics per split, producing a polythetic key in which each splitting point would also have been statistically valid to a level selected by the expert before the run was started.

With some types of data, this methodology applied to automatic key generation would also have offered significant efficiencies in computer use when compared with some existing automatic key construction methodologies, e.g. see Table 19 of this thesis.

In summary, the automatic key-generation option was not pursued because, while automatic key generation may be relevant to the less demanding types of problems encountered elsewhere, taxonomists do not consider automatic key-generation adequate to provide a solution to the types of problems encountered by taxonomists when generating practical keys from data of botanic origin.<sup>3</sup>

---

<sup>1</sup>Pankhurst, Richard J., *Practical taxonomic computing*, Cambridge University Press, Cambridge, 1991, p. 132.

<sup>2</sup>*Ibid.*

<sup>3</sup>E.g. see previous discussion in section 2.2.3 and elsewhere in this thesis.

### 3.2.3 Use of more than one characteristic per decision

Whilst the use of one characteristic is theoretically sufficient for a decision point on a decision key, in practice it is preferable to use more than one characteristic per decision. Polythetic keys are preferred to monothetic keys because the use of multiple characteristics per decision can provide a measure of error-correction.<sup>1</sup> This is examined in section 3.2.3.1. Section 3.2.3.2 suggests ways the use of multiple characteristics could be incorporated in the Selecta-key process.

#### 3.2.3.1 Error Correction using Multiple Characteristics.

Suppose there are a series of  $n$  questions appearing at the branching point of a key. Let two sets of answers to these questions be represented by the vectors  $\mathbf{x} = x_1 \dots x_n$  and  $\mathbf{y} = y_1 \dots y_n$ ; where  $x$  and  $y$  may be binary or non-binary vectors.

The Hamming distance between these two vectors  $\text{dist}(\mathbf{x}, \mathbf{y})$  is defined as the number of places where they differ; for example:<sup>2</sup>

$$\text{dist}(11110, 01010) = 2$$

$$\text{dist}(2201, 2012) = 3$$

The Hamming weight of the vector  $\mathbf{x}$   $\text{wt}(\mathbf{x})$  is defined as the number of non-zero elements contained in the vector  $\mathbf{x}$ , for example:<sup>3</sup>

$$\text{wt}(111010) = 4$$

---

<sup>1</sup>Similarly, 'When there is noise on a channel, however, there is some real advantage in not using a coding process that eliminates the redundancy. For the remaining redundancy helps combat the noise.', Shannon and Weaver, p. 22.

<sup>2</sup>The treatment in this section owes much to portion of the chapter on linear codes in MacWilliams and Sloane, with the difference that the "error correcting" additional characteristics are not produced the same way as the error-correcting extra digits in the linear codes discussed in this reference. MacWilliams, F.J. and Sloane, N.J.A., *The Theory of Error-Correcting Codes*, North-Holland Publishing Company, Amsterdam, 1978. Hamming's work is also discussed in Thompson, Thomas M., *From Error-Correcting Codes Through Sphere Packings to Simple Groups*, The Mathematical Association of America, 1983, pps. 1 - 59. Hamming's original paper was : Hamming, R. W., 'Error Detecting and Error Correcting Codes', in *Bell System Technical Journal*, Volume 26, Number 2, April, 1950, pps. 147-160.

<sup>3</sup>See Wakerly, John, *Error Detecting Codes, Self-Checking Circuits and Applications*, North-Holland, New York, 1978, p. 11; see also Peterson, W. Wesley and Weldon Jr., E.J., *Error Correcting Codes*, The MIT Press, Cambridge, Massachusetts, 1972, p. 40.

$$\mathbf{wt}(21100121) = 6$$

The relationship between  $\mathbf{dist}()$  and  $\mathbf{wt}()$  may be expressed as:

$$\mathbf{dist}(\mathbf{x}, \mathbf{y}) = \mathbf{wt}(\mathbf{x} - \mathbf{y}) \quad (25)$$

Both sides of equation 25 express the number of places where  $\mathbf{x}$  and  $\mathbf{y}$  differ.

Errors in measurement or characteristic identification may occur when vector  $\mathbf{x}$  is evaluated in the light of the available information. Also the factor  $\mathbf{x}_1$  being evaluated may be correctly measured or identified, but if the original separation made by Selecta-key was at the extreme upper level of acceptance of a splitting point (the 5% level of significance) there would be a probability of 0.05 that normal variation in the characteristic would place a specimen on the "wrong" side of the splitting point. Let us define the *error vector* of  $\mathbf{x}$  and  $\mathbf{y}$  as

$$\mathbf{e} = \mathbf{y} - \mathbf{x} = \mathbf{e}_1 \dots \mathbf{e}_n$$

Let the probability of an error occurring be  $p$ . Now:

$$\text{Prob}\{\mathbf{e} = 00000\} = (1 - p)^5$$

$$\text{Prob}\{\mathbf{e} = 01000\} = p(1 - p)^4$$

$$\text{Prob}\{\mathbf{e} = 10010\} = p^2(1 - p)^3$$

and so on...

In general, if  $\mathbf{v}$  is some fixed vector of weight  $\mathbf{w}$ ,

$$\text{Prob}\{\mathbf{e} = \mathbf{v}\} = p^{\mathbf{w}}(1 - p)^{n-\mathbf{w}}$$

Since  $p \leq 0.05$ , then  $(1 - p) > p$  and

$$(1 - p)^5 > p(1 - p)^4 > p^2(1 - p)^3 > \dots$$

It thus follows that a particular error vector of weight 1 is more likely than an error vector of weight 2, and in turn an error vector of weight 2 is more likely than an error vector of weight 3, and so on... This leads to the *minimum-Hamming distance*

strategy which may be used to correct for errors or anomalies when evaluating a multi-characteristic key decision point.

In general, if  $n$  is the number of characteristics employed in a decision regarding a path in the decision key, the number of errors or missing characteristics which can be corrected is  $m$  where:<sup>1</sup>

$$m = \left[ \frac{1}{2}(n - 1) \right]^{\S} \quad (26)$$

If the expert wishes to use ternary or quaternary etc. decisions with multiple characteristics per decision, the use of the minimum Hamming distance will add robustness in these cases as well.

### 3.2.3.2 Multiple Characteristics and Selecta-key

Suppose the evaluation of the elements  $\mathbf{d}_i$  of the decision vector  $\mathbf{d} = \mathbf{d}_1 \dots \mathbf{d}_n$  results in vector  $\mathbf{d}$  being identical to one of the (e.g.) two vectors which would lead to the choice of one path in the decision key.<sup>2</sup> In this case the choice is not in doubt. However if the decision vector is not identical, then the minimum-Hamming distance criteria can be used.

As an example of the application of the minimum-Hamming distance criteria to the Selecta-key methodology, consider the splitting point S3 of Figure 12, and assume that Selecta-key's analysis of the data had shown that four other characteristics had shown the same split as the most favoured characteristic. In this case the situation shown in Table 8 could apply.<sup>3</sup> If the specimen being identified passed questions 1, 3 and 4, and failed questions 2 and 5, the decision vector would be  $\mathbf{d} = (10110)$ . Since this has a Hamming distance of zero from the "pass" vector (10110) and a Hamming distance of 5 from the "fail" vector (01001), the "pass" path is the minimum Hamming distance from the decision

<sup>1</sup>This is also discussed in Hamming, Richard W., *Coding and Information Theory*, Prentice-Hall Inc., New Jersey, 1980, pps. 43 - 49.

<sup>\S</sup>  $\lfloor x \rfloor$  indicates the greatest integer less than or equal to  $x$ ; e.g.  $\lfloor 3.5 \rfloor = 3$ ,  $\lfloor -1.5 \rfloor = -2$ .

<sup>2</sup>This example assumed a two-way, or binary, decision point in the decision tree. The same approach may be used in the case of a multi-way split of 3, 4 or more alternatives.

<sup>3</sup>Table 8 assumes binary vectors, the strategy of using the minimum Hamming distance will apply in the case of vectors with non-binary elements as well.

vector and would thus be the preferred path in the decision key; (this path could then lead to further questions which would lead to the eventual identification of species  $\alpha$  and  $\beta$ ).

	Questions					Path to Species:
	Q1	Q2	Q3	Q4	Q5	
Pass	1	0	1	1	0	$\alpha, \beta$
Fail	0	1	0	0	1	$\gamma, \delta$

Table 8 — Binary decision elements of a multiple characteristic vector.

The case above could have been handled just as well by a question involving a single characteristic; however the advantage of this approach can be seen if specimens with difficult or missing characteristics are encountered.

	Questions				
	Q1	Q2	Q3	Q4	Q5
Specimen A	0	0	1	1	0
Specimen B	0	0	0	0	0
Specimen C	0	?	0	0	1

Table 9 — Binary representation of sample specimens.

The identification of specimens from a botanic data collection often involves attempting to classify specimens into categories which are not easily separable, (e.g. judging leaf shape from a series of printed templates).<sup>1</sup> Use of the minimum Hamming distance for decisions adds robustness to the decision keys as it allows small classification errors to be made without disrupting correct classification. As an example, specimen A of Table 9 would be allocated to the  $\alpha\beta$  choice in Table 8 because it has a Hamming distance of  $\mathbf{dist(10110,00110) = 1}$  from this option, and  $\mathbf{dist(01001,00110) = 4}$  from the  $\gamma\delta$  option. In this case, the

<sup>1</sup>For example, see Figure 1 in Howell, Jim, 'S-Trees: A New Way to Handle Subjective Rules', *AI Expert*, Vol. 7, No. 2, February 1992.

decision process has tolerated a variation in one of the characteristic values without affecting the final decision; i.e. an error correction of one characteristic is possible in this case.

As a second example, specimen B would be allocated to the  $\gamma\delta$  side of the decision key, because  $\text{dist}(10110,00000) = 3$ , and  $\text{dist}(01001,00000) = 2$ . Note that in this case either an error correction of two characteristics has occurred, or the specimen has been wrongly classified because three errors occurred.

Botanic data also very often contains specimens which are incompletely described, due to such factors as the seasonality of many important characteristics. Again in this case, use of the minimum Hamming distance can add robustness to the identification process. As an example, specimen C of Table 9 would be allocated to the  $\gamma\delta$  choice in Table 8 because it has a Hamming distance of  $\text{dist}(01001,01001) = 0$  or  $\text{dist}(01001,00001) = 1$  from this option, and  $\text{dist}(10110,00001) = 4$  or  $\text{dist}(10110,01001) = 5$  from the  $\alpha\beta$  option. In this case the process was robust enough to still allow a valid decision to be made when one of the characteristics was missing. With 5 questions, a valid decision may still be made with 2 characteristics missing.

In these cases, use of only the "best" (Q1) characteristic would have resulted in the wrong classification in the case of specimen A. In the case of specimen C, if the "best" characteristic had been Q2 (not Q1), use of only the "best" characteristic would have resulted in an "unable to classify". Use of the Hamming distance allowed the appropriate decision to be made in each case. This ability is particularly important in the case of botanical specimens, because many of the characteristics which are most helpful in identification are flower and seed characteristics, which are only seasonally available. When these are not available, use of other less indicative characteristics are necessary.

Whether or not the use of multiple characteristics is possible will depend on the form of the data.

The way this will be determined is that, after the best characteristic is chosen (by the method outlined in the previous

section) the expert could again inspect the alternatives offered by Selecta-key. It may well be possible that several other characteristics offer the same species split. These could be added as additional questions to the one already chosen for that key split.<sup>1</sup>

### 3.3 'Voting' Methodology.

An alternative methodology is called the Voting Method. This section briefly discusses the methodology and it's implementation. More detail of the Voting methodology may be found in Appendix C of this thesis, where it is discussed in greater detail than in the following brief section.

When developing the Selecta-key methodology outlined in this chapter, a very much simplified version of the methodology came to mind. It did not appear to have the promise of an accuracy of identification as high as the methodology outlined in this chapter, but it did promise to be a very fast method. Since it is somewhat peripheral to the main thrust of this thesis, it is dealt with in Appendix C. However in outline, with this methodology, the data is split into training and test data.<sup>2</sup> Measurements observed for each characteristic of each species of the training set of data are grouped, and the groups ranked for each characteristic. Splitting points are established for each species per characteristic.

Identification of specimens in the test data can then be made by comparing the measurement for each characteristic with the "template" established from the training data; each species receiving a "vote" if the specimen's characteristic measurement falls within the species' splitting points for that characteristic. The species with the highest "vote" total is declared to be the likely species to which the specimen belongs.<sup>3</sup>

---

<sup>1</sup>For example, in the case of Figure 12, other characteristics which also offered a grouping of ( $\alpha, \beta$ ) on one side of a split, and ( $\gamma, \delta$ ) on the other side could be specified as an additional question for that decision. Characteristics which offered other splits, e.g. ( $\alpha, \gamma$ ) on one side and ( $\beta, \delta$ ) on the other side would be ignored at this stage.

<sup>2</sup>For more detail see section 5.4 of this thesis.

<sup>3</sup>For more detail, see section C.1.1 of Appendix C of this thesis.



### 3.4 Summary — Statistical Methods

Adoption of the methods outlined in this section could assist the identification of botanical specimens, either without a botanical key, or with an understandable botanical key which could be constructed from raw data more easily than has been the case in the past.

While some of these processes would be computationally intensive, they provide a statistically sound method of ordering questions. The expert can then choose from the options offered by Selecta-key question(s) which are practical for the application concerned. It is envisaged that this combination of (computationally intensive) statistical validity and human common sense would best use the strengths of both computer and human expert to provide a useable and useful key for later identification of specimens of the species or taxa being examined.

# INDUCTIVE CATEGORISATION IMPLEMENTATION

This chapter discusses implementations of the inductive categorisation algorithms introduced in the previous chapter. The implementation of these inductive categorisation algorithms will be referred to in the rest of this thesis as the Selecta-key programs. This chapter also contains some brief comments on some other necessary programs developed during the course of this study to supplement the Selecta-key programs.

Sections 4.1, 4.2 and 4.3 discuss the first implementation, second implementation and third implementation respectively of Selecta-key prototypes, and some of the limitations imposed by practical considerations.

Section 4.4 deals with the implementation of the simplified identification methodology derived from the Selecta-key methodology, the voting method.

Section 4.5 outlines a program used for detecting outliers and indicating possible data errors in the data used by Selecta-key and the other classification programs.

Section 4.6 discusses the implementation of a neural net designed to handle input which is divided into categories. It also notes that a neural net able to handle real number data was not developed because a versatile simulator became available. The neural nets were used to compare classification accuracy with the Selecta-key methodology.

Section 4.7 comments briefly on the implementation of ancillary data conversion and other programs necessary to allow the results obtained by Selecta-key runs to be compared with the results obtained by the use of other methodologies.

## 4.1 First prototype of Selecta-key

The first prototype implementation of the Selecta-key methodology employed the small-sample student's *t* test to test the null hypothesis that there was no difference between the

distributions being examined. It was run with a wide variety of limited, artificial test data, producing keys in a screen format similar to those appearing in Collier's paper.<sup>1</sup> The data space requirements of this prototype meant that a full set of *Acaena* data was too large for the implementation.<sup>2</sup> The initial results had to be obtained from a sub-set of the *Acaena* data.<sup>3</sup> The sub-set was chosen from *Acaena* specimens which had particularly complete data, and was thus unlikely to be truly representative of data likely to be encountered in the field.

The first prototype also had other limitations. It only allowed the use of a t test. It thus did not allow the use of either normal or randomisation tests, and the program style employed did not allow the program to be easily extended to allow the inclusion of these tests (without running into further capacity problems). The prototype was initially implemented in Turbo Pascal 3.0 on a Unitron 2900S, and memory limitations inherent in the configuration made careful re-writing necessary if suitably large data sets were to be handled.<sup>4</sup>

Despite these limitations, the first prototype did all that was expected of it, and provided an excellent background for the next prototype.

## 4.2 Second prototype of Selecta-key

The second implementation added normal distribution tests to the t test implemented in the first prototype. These allowed examination of the null hypothesis that there was no difference between the groups of samples (i.e. that they were random selections from the same normal distribution). Compared with

---

<sup>1</sup>P.A. Collier, *Inductive Inference for Botanical Keys*, in Proceedings of the Third Australian Conference on Applications of Expert Systems, The New South Wales Institute of Technology, Sydney, 1987.

<sup>2</sup>The full set of *Acaena* data was also too large for the commercial inductive categorisation program *1st Class*, running on the same IBM-PC clone, with which the results from the Selecta-key methodology were being compared.

<sup>3</sup>The background to the choice of the *Acaena* data as a data set used for comparative runs between different methodologies is given in the next chapter of this thesis.

<sup>4</sup>The Unitron 2900S was an 4.77 MHz, 8088 IBM-PC compatible running PC DOS 3.01, initially with 256K memory, two 360K floppy drives and a colour CGA screen. During the development of the Selecta-key programs, the memory was enlarged in stages to 640K, an 8087 hardware multiply chip and then a 10 megabyte hard disk were added.

the t test, the normal distribution tests had the advantage that separate standard deviations could be allocated to each group, and hence better splitting points were obtainable. However concomitant with this was the disadvantage that more specimens were needed per characteristic per species to satisfy the usual requirements for a normal distribution.<sup>1</sup>

The second prototype was written in such a way that, (unlike the first prototype), it could be easily extended. Features implemented during the life of this prototype included the ability for only some species to be separated in a split, (one or more distributions being able to be placed in each of the resultant separate groups after a split). As an example of this, consider the case of Figure 19, a split could be made between distributions  $\alpha$  and  $\delta$ , with distribution  $\beta$  being allocated to both sides of the key; the separation between  $\alpha$  and  $\beta$ , and  $\beta$  and  $\delta$  being made at the next stage of the key. This feature was added because, despite indications by some theoretical approaches that this practice would produce a sub-optimal key, it was found to be quite useful in dealing with some difficult data; a more complete key being able to be produced.

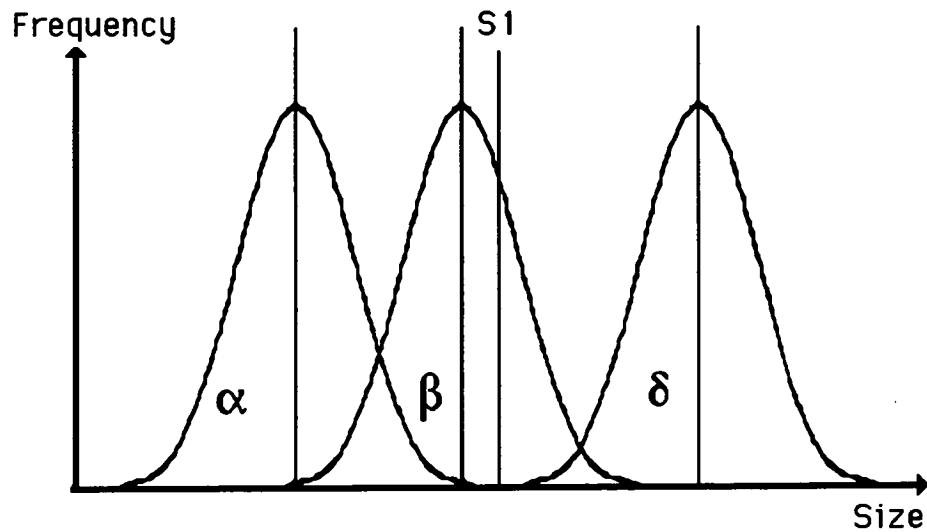


Figure 19.  $x_{split}$  in a distribution

This prototype allowed the level of rejection of the null hypothesis (used in the t and normal tests) to be fed in as data. During a run it presented the choices at each level of the key ordered according to level of confidence of a split with that

<sup>1</sup>See section 3.1.1 of this thesis, second-last paragraph.

characteristic, allowing the user to select the characteristic to be used as the splitting characteristic at that level of the key; it allowed the user to mark some species as unseparable, and produced a key diagram after the run was complete.

The main disadvantage of the second prototype was that, despite careful programming and enlarging the memory of the IBM-PC compatible, a run using the *Acaena* data was only just able to be handled, and the *Danthonia* data was still too large.<sup>1</sup> In this case, adding the ability to handle randomisation tests seemed, in practice, out of the question on this platform.<sup>2</sup>

### 4.3 Third prototype of Selecta-key

Capacity and other problems led to successive Selecta-key implementations on a Macintosh Plus and SE, Prime 9955, Sun 4, and an attempt to implement the package on an IBM RS6000.

Whereas the language used for the first and second prototypes was Turbo Pascal 3 and 4, versions of the third prototype started life under Turbo Pascal 4 & 5.5, and were later converted to Think (LightSpeed) Pascal when this became available. Subsequently further conversions took place using Prime Pascal, Sun Pascal 1 & 2, and RS6000 Pascal.<sup>3</sup> Constant conversions became wearisome, and a set of transportable

---

<sup>1</sup>An explanation of the reason that these data sets were chosen for use in comparative runs between different methodologies is given in the next chapter of this thesis.

<sup>2</sup>The Selecta-key implementation was written in Turbo Pascal 3.0 and 4.0, when the latter became available. By this time, the implementation had been split into several parts, each of the parts being chained together automatically during a run. The largest array had been converted to a disk array, but the use of more than one disk array seemed difficult. Overlaying was investigated as an additional strategy, but was judged not to offer sufficient potential capacity. Despite the use of an efficient B-tree data storage scheme, the use of all these strategies meant that a run was by now quite slow; hence it seemed that a change of platform was appropriate.

<sup>3</sup>The change to Think Pascal on the Macintosh was made to escape PC DOS 3.10's 640K memory limit. The change to Prime Pascal was made to gain more speed, and to escape the small memory sizes on the available Macintoshes when expected memory upgrades did not eventuate. The change to Sun Pascal 1 was made because the Prime was due to be de-commissioned, and proved a good choice, as Sun Pascal proved much more capable and reliable than Prime Pascal. A change to the IBM RS6000 was attempted because the Computer Science Department was attempting to centralise all its computing on the newly-purchased machine (not because of any problems with the Sun implementation). However the IBM RS6000 Pascal's implementation of sets proved too fragile to allow reproducible, consistent runs to be obtained at that time. Work reverted to the Sun 4/Pascal 2 combination, which proved reliable. Work has continued on the Sun 4 since.

utilities were written which allowed the same Pascal programs to be run on the different machines with only minor alterations.<sup>1</sup>

Implementation of the third prototype added randomisation tests and polychotomous splits to the normal and t tests implemented in the first prototype.

Randomisation tests allowed examination of the null hypothesis that there was no difference between the groups of samples (i.e. that they were random selections from the same distribution) regardless of the form of the distribution.

The existence of polychotomous decisions in any particular set of data is not certain until the program is run, however they may be used if the data can support them. Often, however, the lack of separation between the distributions is such that it may prove preferable in practice to only use some of the available splitting points at any decision point in the key, even if more are available. Dichotomous or polychotomous splits are only accepted if the null hypothesis (that any distribution below the splitting point is drawn from the same distribution as any distribution above the splitting point) can be rejected at a pre-determined

---

<sup>1</sup>Because of the inadequacies of the Pascal standard, there are considerable differences between different implementations of Pascal on different platforms. These differences can cause considerable problems when transferring programs between different platforms, particularly if the code has been optimised for a particular implementation, (in some heavily optimised cases a major re-write may even be required). The compatibility packages were developed in an attempt to hide many of these differences in implementations. The package for an individual platform was usually written as a series of units (or modules, or whatever the local equivalents were called). The package usually consisted of a series of units which contained procedures which hid or made allowance for fundamental Pascal implementation differences such as ascii-ebcdic character representation, reading and writing files, string manipulation, real number accuracy (this partially), maximum integer size etc.; but specifically excluding the local equivalent of the 'include' statement for use with units. Both the 'include' statement and units themselves are non-standard, but an equivalent exists in virtually all industrial-strength implementations of Pascal. The procedures also offered a few desirable extras, such as text data files which could contain clarifying comments which would be ignored on subsequent data input, the equivalent of command-line variables, and several standardised forms of enquiry to the user (e.g. allowing the specification of restricted characters responses (if required), default responses, the ability to specify integer instead of real responses and to be warned if this was not adhered to, the ability to recover cleanly if foreign characters (e.g. letters) were entered as part of a number, question-specific help, question-specific error messages, the ability to halt cleanly if the user requires this, etc.) to be presented at every question using this system. When a program had been written with these compatibility procedures in mind, all that was required to change to another machine was to change the program's 'include' statement to the local equivalent. Implementing these procedures took a considerable amount of time and thought, but has proven very useful in the long run.

level of confidence. In this prototype (as in the others) this level is read in as data, as follows:<sup>1</sup>

```
franklin% select
```

The program running is Select.p version 0.04. This program uses a research version of Selecta-key's methods to help produce a key.

Do you need more detail about this program? <n>

Do you wish to use this program? <y>

Run commenced at 5:17 a.m. on Friday 4-Dec-1992

Setting up...

The null hypothesis states that the "two" samples are in fact drawn from the same population, and any difference between them is merely an artifact of chance.

Level to reject null hypothesis [e.g. 5%]=<5.0>

Number of standard deviations to accept split [e.g. 2] =<2.0>

Which data:-

a=acaenac, u=acaenau, o=golforig, e = weedseed, g=golfmod

w = wornseal, s=side, t=table, x=other file; (a/u/o/e/w/s/t/g/x) ?<e>x

Please type in the name of the file you wish to use:<golf.sk> acD80.sk

The program gives the user the opportunity to specify that all splits be on a statistical basis, all be on a randomisation basis, or the individual tests may be specified as either one or the other:

Do you wish all the test to be "random" tests (r), or

"statistical" tests (s), or

your choice of "random" and "statistical" mixed, (m): <s>

Species to be examined are as follows:-

1 esub	2 eret	3 eech	4 erob
5 etyl	6 aagn	7 aten	8 aaeq
9 apro	10 oovi	11 ovel	

Preparing data for split point calculation:-

If all data is missing from a characteristic of a species, a warning is printed:

WARNING:-

The numeric characteristic LenShortSp  
of species erob  
has no legitimate data.

WARNING:-

The numeric characteristic LenShortSp  
of species apro  
has no legitimate data.

---

<sup>1</sup>In the following fragment of a program run, it will be noted that use of the transportable utilities allows the use of default answers enclosed in angle brackets <> to most of the questions. In most cases below the default values are accepted by pressing the *Return* key. *franklin* is the name of the computer on which the program is run.

A sample output from the program when a single split is possible is:<sup>1</sup>

Using the acD80.sk data, species with low numeric values listed first...  
Characteristic being used for this split = PetioleHo

aten	apro	oovi	aagn	aaeq	ovel
Null hypothesis rejection level, (worst)= 1%, split value =1.8481					
etyl	eech	esub	eret	erob	

Do you wish to accept this split (a), examine the next (n) or  
previous (p) one, decide the species are indistinguishable (i)  
or collect more information about a split (c)? (a/n/p/i/c)<n> c

If the user is uncertain about a split, a diagnostic output patterned on Figure 15 of this thesis is optionally available:<sup>2</sup>

```
Characteristic = PetioleHo
1 aten      1.00| **
2 apro      1.00| 00 **
3 oovi      1.25| 13 13 **
4 aagn      1.33| 16 16 77 **
5 aaeq      1.35| 0 0 57 93 **
6 ovel      1.46| 1 1 39 67 60 **
7 etyl      2.45| 0 0 0 0 0 1 **
8 eech      2.50| 0 0 0 0 0 0 90 **
9 esub      2.59| 0 0 0 0 0 0 69 69 **
10 eret     2.83| 0 0 0 0 0 0 25 10 11 **
11 erob     3.00| 0 0 0 0 0 0 8 1 0 4 **
               |=====
               | 1 2 3 4 5 6 7 8 9 10 11
```

If the user decides to accept this split, then next (n) or previous (p) outputs may be examined to see if there are any other characteristics which give the same split (to get the advantage of an error-correction effect). A suitable additional characteristic to use at that split would be:

Using the acD80.sk data, species with low numeric values listed first...  
Characteristic being used for this split = HbLeaflet

esub	eech	erob	eret	etyl
Null hypothesis rejection level, (worst)= 0%, split value =3.4152				
ovel	aaeq	oovi	aagn	aten
				apro

Do you wish to accept this split (a), examine the next (n) or  
previous (p) one, decide the species are indistinguishable (i)  
or collect more information about a split (c)? (a/n/p/i/c)<n> c

<sup>1</sup>The maximum length of the species and characteristic names are limited by some of the implementations used to compare the Selecta-key results. Use of computer-translated data, while it had the advantage that it guaranteed that the data was comparable for each alternative methodology being tested, also had the disadvantage that it led to the lowest common denominator of species and character description lengths being used. These descriptor length limits are not an inherent part of the Selecta-key implementation.

<sup>2</sup>The table entry '00' represents 100%. The numbers to the left of the Table are mean values for that characteristic/species.



If one of the characteristics has no valid values, the following type of diagnostic output is available:

```
Calculating split points:-
Characteristic = LenShortSp
 1 erob      ???| **
 2 apro      ???| 00 **
 3 oovi      1.15| 00 00 **
 4 etyl      1.20| 00 00 90 **
 5 esub      1.35| 00 00 11 71 **
 6 aagn      1.40| 00 00 42 69 87 **
 7 eech      1.51| 00 00  2 46 31 73 **
 8 ovel      1.54| 00 00  4 44 33 68 90 **
 9 eret      1.78| 00 00  0 15  0 20  4 16 **
10 aaeq      1.82| 00 00  0 13  0 18  5 14 72 **
11 aten      1.97| 00 00  0  6  0  7  0  3  6 22 **
      |=====
      | 1  2  3  4  5  6  7  8  9 10 11
```

As in the first two prototypes, the options available for choice as splitting points are displayed. In this case however, the strength of separation of each distribution from the splitting point can be displayed. The minimum strength is noted for each characteristic, and all the potential splitting points are presented to the expert in an order to allow the selection of the characteristic(s) considered the most appropriate by the expert. It will be noted in the program output fragment below that one split occurs at the 4% level; even though it is a legitimate split the user may prefer to ignore this:<sup>1</sup>

Using the acD80.sk data, species with low numeric values listed first...  
Characteristic being used for this split = Stamens

```
ovel      aaeq      oovi      eech      aten      esub
eret
Null hypothesis rejection level, (worst)= 0%, split value =4.7444
etyl      aagn
Null hypothesis rejection level, (worst)= 4%, split value =5.9714
apro
Null hypothesis rejection level, (worst)= 0%, split value =6.5000
erob
```

```
Do you wish to accept this split (a), examine the next (n) or
previous (p) one, decide the species are indistinguishable (i)
or collect more information about a split (c)? (a/n/p/i/c)<n> c
```

If the data is such that some species are indistinguishable, they may be marked as such, and construction of the key may then continue. If no split is possible, an output of the type below is seen:

---

<sup>1</sup>This is done by choosing the 'c' option. After printing out some diagnostic information the program then gives the user an opportunity to enter their own preferred split value.

Using the acD80.sk data, species with low numeric values listed first...  
Characteristic being used for this split = LenShortSp

No Species could be distinguished via this numeric characteristic.

Do you wish to accept this split (a), examine the next (n) or  
previous (p) one, decide the species are indistinguishable (i)  
or collect more information about a split (c)? (a/n/p/i/c)<n>

The version of the Selecta-key methodology proved to have the ability to handle adequately large data sets, for example the *Danthonia* data could now be handled. More extensive tests were carried out with this version, producing satisfactory results which are presented later in this thesis.

## 4.4 Simplified Inductive Classification (Voting).

This methodology arose as a simplification of the Selecta-key methodology. The methodology is outlined in section 3.3 of the previous chapter, and explained more fully in Appendix C.

This methodology was implemented in Pascal 2.0 on a Sun 4 computer, using the portability package developed as a part of this project. As implemented, the programs only handle cases where the data is complete, and hence results are only presented for the *Danthonia* data.<sup>1</sup> The program uses the same input format as the Selecta-key programs.

## 4.5 Checking for Outliers

Some of the tests employed in this thesis are quite sensitive to the presence of outliers in the data.

The presence of multiple outliers in a specimen's measurements may be indicative of a mis-classification or other data error.

To fulfil the need to quantify the magnitude and frequency of outliers in the data, a program was written which would examine the data for the presence of outliers. Any specimen characteristic of any species which was more than two standard

---

<sup>1</sup>Extension of this methodology to include the ability to handle data with missing values is possible, subject to the precautions mentioned in section C.1.2 of Appendix C of this thesis.

deviations from the mean of that characteristic's measurements was reported by this program.

This program was implemented in Pascal 2.0 on a Sun 4 computer, using the portability package developed as a part of this project. As implemented, the program treats categorised data as having an order imposed by the data conversion program mentioned as b) in section 4.7 of this chapter. Outliers reported for categoric data should be treated with caution, and the ability of program b) of section 4.7 to allow the user to change these allocated numbers should be remembered. The program uses the same input format as the Selecta-key programs. The results obtained by the use of this program are discussed in section E.4.2 of Appendix E.

## 4.6 Aristotelian Neural Net Simulator.

It was considered desirable to compare the Selecta-key approach with the (at the time) relatively newly revived neural net approach. Since no simulators were available to the author, it was planned to implement two neural net simulators. The first would handle categoric input (of the type used in some early experiments on pattern recognition). The second would handle real number input, and have an ability to generalise which was hoped to be superior to that expected to be exhibited by the first implementation.

Of the types of neural net which could have been implemented, the multi-layer perceptron net was chosen.<sup>1</sup>

Firstly a neural net simulator was developed which adhered to Aristotle's rules of inductive logic, i.e. it assumed complete enumeration, reporting an error if this assumption was found to be invalid. This was developed in Turbo Pascal 4.0 on an IBM-PC clone, and later transferred to the Sun 4.

The second neural net simulator was approximately three-quarters developed when the MITRE simulator became available.<sup>2</sup> The MITRE neural net simulator is a versatile

---

<sup>1</sup>The reasons for this choice are discussed in Appendix B of this thesis.

<sup>2</sup>See Leighton, R., and Wieland, A., *The Aspirin/MIGRAINES Software Tools User's Manual, Release 4.0*, The MITRE Corporation, Washington, 1991.

simulator which can run on a variety of machines, including the Sun 4. Since there seemed little point in re-inventing the wheel, development of the second neural net simulator was stopped, and the MITRE simulator used for comparisons with the Selecta-key methodology.

## 4.7 Ancillary programs

To assist comparison of the results obtained by the Selecta-key and other methodologies, programs were written to convert data into formats appropriate for the available implementations of the different methodologies.

A data format similar to that originally used by Collier's ID3 implementation was taken as the reference standard. The format was the same as Collier's original ID3 input format, with the exception that, in the standard format, specimens of the same species were grouped together in the data.

In the course of this work data conversion programs were written which translated data:-

- a) from the original Selecta-key input format to the reference format;<sup>1</sup>
- b) from the reference format to the Selecta-key input format, (allowing the selection of various means of splitting the data into sub-sets of the data, and the re-ordering and/or combining of categoric data). This program also reported any species not represented in either of the training or test data sets. This was necessary because several of the implementations of the comparison methodologies will not produce a result if this condition occurs);<sup>2</sup>
- c) from Selecta-key input format to Collier's original ID3 implementation input format;<sup>3</sup>

---

<sup>1</sup>Necessary to change the original versions 1 and 2 of the Selecta-key prototype's data files into reference format, and hence allow them to be split up into training and test data files to be used with the third prototype.

<sup>2</sup>Used for splitting the reference data files into sets of training and test data files (for more detail see section 5.4 of this thesis). It also provided a means of testing to see if there was any effect caused by different ordering of the categoric data.

<sup>3</sup>Necessary to allow comparison runs with Collier's original ID3 implementation.

- d) from the Selecta-key input format to two files which are in Quinlan's standard data format. This program also optionally allows the output of the data in a file format which is acceptable as input to an Excel spreadsheet;<sup>1</sup>
- e) from Selecta-key input format to a format suitable for input into the Aristotelian Neural Net simulator, (note that this program also categorises any real number data during this conversion). This Neural Net simulator was not able to conveniently handle missing data, so the data conversion program noted the range of the particular characteristic, and randomly allocated a data value within that range if the data value was missing in the case of that particular specimen. To prevent this "synthetic" data interfering unduly with the training or testing, a facility was added to this data translation program to allow multiple copies of the data to be included in the translated data, the "real" data being the same in each copy of the data, but the "synthetic" data being different random values (each within the noted appropriate range) in each version of the data. The number of multiple versions (the multiplication factor) could be specified by the user;
- f) from Selecta-key input format to a format suitable for input into the MITRE Neural Net simulator. The MITRE simulator was also not able to conveniently handle missing data, so the ability to generate "synthetic" data was also built into this program, in the same way as outlined in the paragraph above;
- g) from Selecta-key input format to a format suitable for input into the SAS statistics package.<sup>2</sup> This program also produces a batch file suitable for use in running the application in the background, detached from the terminal.

---

<sup>1</sup>Part way through this process, Collier changed the input format of his TL implementation of ID3 so that it matched the data format used by J. Ross Quinlan's programs. This data conversion program was necessary to allow continued comparison runs using Collier's ID3 implementation. The Excel data format output file was a hangover from some early investigatory work using this spreadsheet.

<sup>2</sup>SAS is a registered trademark of SAS Institute Inc..

Use of these programs ensured that the data used for comparative runs was, for the purposes of calculation, identical to that used in the Selecta-key runs.<sup>1</sup>

Other ancillary programs also had to be written to allow the work to progress.<sup>2</sup>

---

<sup>1</sup>For more information on the data requirements, see the following chapter of this thesis. There are references in this chapter to the more detailed examination of the data which is carried out in Appendix E and part of Appendix A.

<sup>2</sup>Since the network was new, it was also necessary to write high level drivers for three printers, which allowed:-

- a) sending text and postscript files through the network to a specified laserwriter. The driver for the text files gave a choice of font type, size, page orientation and number of columns per page, and corrected the number of lines per page for some configurations;
- b) sending text files through the network to a specified ink jet printer, giving a choice of font type, size, and page orientation. It also corrected the (incorrect) default options for the number of characters per line and number of lines per page in each of these options;
- c) sending text files through the network to a specified wide-paper continuous feed dot matrix printer.

All of these were implemented in Pascal 1 on a Sun 4. The Pascal 1 programs required the incorporation of some C code. Most have since been converted to Pascal 2.

# Inductive Categorisation, Dendrograms, and Botanical Data

An examination of the literature on the construction of dendritic trees (dendrograms or keys) in botany revealed that there were three types of dendrograms commonly used in botanic literature, only one of which was able to be constructed by the approach outlined in this thesis. The three types, and the applicability of the Selecta-key methodology in their construction, are discussed in section 5.1. To help distinguish the types of dendrograms, the identification dendrograms will be mainly referred to as keys in the remainder of this thesis.

To establish the relative effectiveness of the Selecta-key versus other methodologies in the construction of botanic keys, comparative tests were necessary. To do this, it was considered important to obtain data sets which contained the types of data problems which can be endemic in sets of measurements obtained from collections of botanic specimens. The types of problems which can typically occur in collection of botanic data are discussed in section 5.2. The suitability of the data sets chosen are then discussed in the light of these requirements, see section 5.3.

To test many of the key construction methodologies it was necessary for the data be split into training and test sets. The methodology employed for this purpose is discussed in section 5.4.

Section 5.5 summarises the findings of this chapter.

## 5.1 Botanic Dendrograms — Limitations of the Selecta-key approach.

In botanical and biological work there have been three traditional uses for dendrograms; genealogical, cladistic and identification dendrograms. It is not the purpose of this section to discuss their merits and thus enter into what Dawkins has referred to as 'one of the most acrimoniously controversial fields in

all of biology'.<sup>1</sup> This section aims merely to delineate the three types, and to emphasise that the Selecta-key methodology is limited in that (while an expert could use the methodology to produce dendrograms of either of the other two types) the methodology only inherently produces dendrograms of one of these types. Section 5.1.1 looks briefly at genealogical dendrograms. Section 5.1.2 examines cladistic dendrograms. Section 5.1.3 notes that the Selecta-key methodology would be mainly used for assisting in the production of identification dendrograms. Identification dendrograms will be mainly referred to as *keys* in the remainder of this thesis, to help distinguish them from these other types of dendrograms used in botanical work.

### 5.1.1 Dendrogram types — Genealogical Dendrograms

Darwin suggested 'our classifications will come to be, as far as they can be so made, genealogies; and will then truly give what may be called the plan of creation.'<sup>2</sup> Dunn and Everitt concur, giving a similar (but less theological) definition in which a dendrogram is 'an evolutionary tree [which is] a summary of a scientific theory to be tested by further research', indicating that sharing a position in the branches of the dendrogram means sharing a common ancestor.<sup>3</sup> These trees were also referred to as cladograms (after Hennig's work) which became widely used, perhaps because 'A unique feature of Hennig's concepts is that they are largely understandable'<sup>4</sup>

---

<sup>1</sup>Dawkins, Richard, *The Blind Watchmaker*, Longman Scientific & Technical, Harlow, England, 1986, p. 255.

<sup>2</sup>Darwin, Charles, *Origin of Species*, 1859 (not seen), quoted in Humphries, Christopher J. and Parenti, Lynne R., *Cladistic Biogeography*, Clarendon Press, Oxford, 1989, p. 22. No fuller reference or page number to Darwin is given by Humphries and Parenti.

<sup>3</sup>Dunn, G., and Everitt, B. S., *An introduction to mathematical taxonomy*, Cambridge University Press, Cambridge, 1982, p. 122.

<sup>4</sup>Nelson, Gareth and Platnick, Norman, *Systematics and Biogeography, Cladistics and Vicariance*, Columbia University Press, New York, 1981, p. 139.



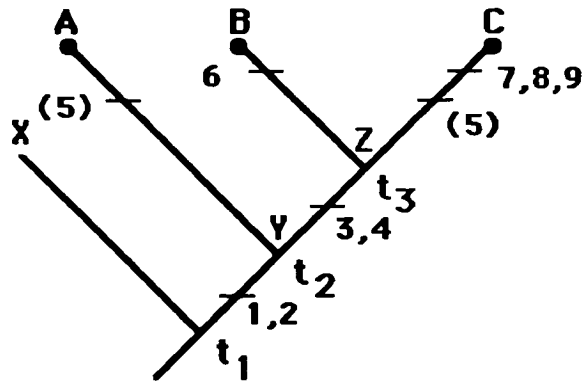


Figure 20 A cladogram showing Hennig's definition of a relationship.<sup>1</sup>

An example is shown in Figure 20. In this case A, B, C and X represent modern taxa. Y and Z represent ancestral taxa.  $t_1$ ,  $t_2$ , and  $t_3$  represent time intervals. The numbers 1 to 9 represent characteristics of the specimens. In this case taxa A, B and C represent a monophyletic group, as they have a common ancestor Y. X is not in this group as it does not share the ancestor. Characteristics 1 and 2 are common to each of the taxa A, B and C. Hennig called these shared derived characteristics synapomorphies if they were inherited from the most recent common ancestor, and simplesiomorphies if they were inherited from a more remote common ancestor. In this case characteristics 1 and 2 would be synapomorphies to taxon A, but simplesiomorphies to taxa B and C. Taxa B and C are considered separate from taxon A because B and C share characteristics 3 and 4 which are unique to this group. Taxa B and C are considered to have a common ancestor Z, but are represented as separate taxa because C exhibits unique characteristics 7, 8 and 9 which do not appear in specimens of taxon B. The characteristics unique to the group are called autapomorphies by Hennig. B also exhibits characteristic 6 which does not appear in taxon C, reinforcing the distinction between these taxa. If a characteristic is similar, but is considered to have arisen separately (e.g. wings being common to both bats and birds) it is represented in brackets, e.g. see characteristic 5.<sup>2</sup>

<sup>1</sup>Similar to figure 2.2 in Humphries and Parenti, p. 23. They reference Hennig, W., *Phylogenetic systematics*, University of Illinois Press, Urbana, U.S.A., 1966 (not seen).

<sup>2</sup>These concepts are also discussed in Nelson & Platnick, Chapter 3, pps. 63 - 168.

The keys produced by Selecta-key are not inherently representative of this type of traditional botanic key usage.

### 5.1.2 Dendrogram types — Cladistic Dendrograms

A second use of dendrograms occurs in cladistic taxonomy, in which specimens are grouped with others who share similar biochemical, morphological or other traits, 'the ultimate criterion for grouping organisms together is closeness of cousinship'.<sup>1</sup> The practitioners of numerical taxonomy fall into this group, and Dawkins comments that some are called 'pheneticists'.<sup>2</sup> In this case no particular account is taken of time intervals.<sup>3</sup>

The keys produced by Selecta-key are not inherently representative of this type of traditional botanic key usage.

### 5.1.3 Dendrogram types — Identification Dendrograms

The dendrograms produced by the Selecta-key methodology discussed here are of a third type, in which the dendrograms 'act as identification keys based on similarities or dissimilarities between different groups.'<sup>4</sup>

The characteristics used by Selecta-key may be synapomorphies, simplesiomorphies or autapomorphies. The

---

<sup>1</sup>Dawkins, Richard, *The Blind Watchmaker*, Longman Scientific & Technical, Harlow, England, 1986, p. 258. For an example of a "traditional" approach to this subject, see Boyd, William C., 'Modern Ideas on Race, in the Light of Our Knowledge of Blood Groups and Other Characters with Known Mode of Inheritance', in Leone, Charles A. (Ed), *Taxonomic Biochemistry and Serology*, The Roland Press Company, New York, 1964, pps. 119 - 169. However it is worth noting that since these books were published, the field which produces both genealogical and cladistic dendrograms has been in a ferment with different results being obtained from the morphological, biochemical and karyological fields. 'As an example, Avise cites the New World and Old World vultures, which share behavioural (soaring) and anatomical (bald head and face) characteristics. Based on these [phenotypical or morphological] features, these vultures were long considered to be close evolutionary cousins. But the [genotypical or biochemical] technique of DNA-DNA hybridisation ... has recently shown that the two birds are only distantly related. ... Not surprisingly, many traditional anatomists were irked to think of themselves as being made redundant by the advent of an alien form of analysis, and they protested loudly and bitterly.' 'The comparison between morphological and molecular evidence too often degenerated into a "Which of these data bases is superior" shouting match'. Both quotations are from Lewin, Roger, "Scenes from a biological revolution", *New Scientist*, 5 March 1994, New Science Publications, London, 1994, p. 42-43.

<sup>2</sup>Dawkins, p. 279.

<sup>3</sup>Ridley examines the basis of these two approaches (and their variants) in some detail, and argues strongly for the former; see: Ridley, Mark, *Evolution and Classification, The Reformation of Cladism*, Longman, London, 1986.

<sup>4</sup>Humphries and Parenti, p. 22.

methodology makes no attempt to distinguish between the characteristics used on the basis of ancestry. Both the *Acaena* and *Danthonia* data used in this thesis may be considered monophyletic groups, but this is not necessary for the Selecta-key methodology to be applied. The methodology will work just as well for data which has no botanical or zoological similarity, e.g. it has been used to examine meteorological data which is associated with the prediction of storms.<sup>1</sup>

An expert may use and/or be assisted by the methods discussed in this thesis to produce dendrograms of the first two types mentioned. Use of the methodology may help distinguish between synapomorphies and simplesiomorphies. The methodology can produce multi-way decisions for characteristics which are not autapomorphies, i.e. where data exists for all taxa of the characteristic being examined. It is also applicable if autapomorphies do exist, but the identification rate may (but not invariably) be lower, as the characteristic may be categoric (e.g. present/not present).

However it should be stressed that production of keys claiming to exhibit the property of either of the types of dendrograms mentioned in sections 5.1.1 and 5.1.2 would depend on and have to be justified by the expert producing the dendrograms, as the ability to automatically produce these types of dendrogram is not an inherent property of the methodology described in this thesis.

## 5.2 Requirements of Botanic Data

The proposed Selecta-key methodology will only be justified if its employment allows the production of paper keys whose use provides an accuracy of identification roughly comparable with existing computer-based methodologies. To allow a realistic comparison of alternative methodologies for the identification of botanic specimens, suitably typical botanic data is essential if bias due to the use of data which conceals or omits the types of difficulties often found in data of botanic origin is to be avoided.

---

<sup>1</sup> Since this thesis is primarily concerned with the application of the Selecta-key system to the construction of botanic keys, the only mention of the storm work in this thesis is in Table 19 in section 6.3 of this thesis.

Requirements for this data for the purposes of methodology comparison are discussed in section 5.2.1.

If it is a requirement that, in addition to comparing methodologies, an accurate and useful key is to result from the processes, then further requirements are placed on the data. These requirements are discussed in section 5.2.2.

### 5.2.1 Data Requirements for Comparison of Methodologies for the Identification of Botanic Specimens

Requirements for data which is to be used for the purpose of methodology comparison fall into two main categories.

Firstly, the data should be representative of the type of botanic data which would be submitted to the methodology if that methodology were to be used in practical situations.

Secondly, the form of the data should encompass the types of problems which routinely occur in data of botanic origin.

To satisfy the first requirement, it would be ideal if the data consisted of measurements and/or classifications of real botanic specimens, rather than constructed data. There is a possibility that constructed data could, perhaps inadvertently, favour a particular methodology in a way that is not typical of the type of botanic data that would be met in the practical application of these methodologies.

To satisfy the second criteria, a fairly detailed examination of the data proposed to be used would be necessary.

The data should be examined to see if the data fits the assumption that it is of a parametric form, e.g. that it could be accepted as likely to be data which fitted the normal or Gaussian assumption. Since botanic data can contain both parametric and non-parametric data, the data used to test the methodologies should contain examples of both types of distributions. This knowledge is also important because some of the methodologies make an assumption that the data belongs to a certain type of distribution, e.g. a normal distribution in the case of multi-variate normal discriminant analysis.

Botanic data can be prone to contain outliers, e.g. it may not be unusual for the height of mature grass to vary by, say, an order of magnitude (depending on shade and soil condition); whereas this type of variation would be unusual in (e.g.) the height measurements of mature human subjects used in psychological observations. It is thus important that the botanic data sets to be used in comparing the methodologies be shown to contain outliers, because some methodologies are more affected by outliers than others. As an example, the t test is relatively insensitive to the form of a distribution, but is affected by the presence of outliers. The choice of clustering methodologies would also be effected by the presence of outliers.

Some methodologies (including the t test) advise that the groups of observations be statistically independent. Unless the characteristics are carefully chosen, this requirement is often not met in the case of botanic data. It would thus seem advisable that the correlations between the observed characteristics be checked to ensure that they are not so high as to invalidate the methodologies being compared.

Often the seasonal nature of some characteristics (e.g. flowers and seeds) means that observations of these characteristics may not be obtainable for all specimens, as some species may grow in geographically disparate locations and travel time and financial considerations may preclude complete data collection. The botanic data used to compare the methodologies should therefore contain some specimens which do not have observations recorded for every characteristic, i.e. there are some measurements missing from the data. Artificial data could be prepared by randomly omitting measurements from complete data, but this is unlikely to be representative of data gathered from the field because the missing data would be randomly selected. In the case of missing botanic data, it is not randomly selected data that is missing, it often occurs that it is *the most important data* in providing identification that is missing. Flower and seed data are very often vital in specimen identification, but it is precisely these observations which may be missing for the reasons noted above. Again, for these reasons, data obtained by measuring and/or classifying real specimens is to be preferred over artificial data.

If the botanic species is a rare one, there may be difficulty in obtaining a statistically sufficient number of observations. It would thus be preferable that the botanic data used to test compare the methodologies also reflect this real-life problem in at least some portion of it's data.

Botanical specimens are classified into species or taxa by processes which are essentially phenetic; i.e. species or taxa are classified into groups based on the similarities or dissimilarities between different groups. Sneath refers to this as isological classification.<sup>1</sup> Thus the botanical species or taxa (for which a key is to constructed) have been classified as being of the same taxa by being noted to be more similar on observable characteristics than other plants. The fact that, when attempting to build a botanic key, an attempt is being made to separate species or taxa which have been deliberately chosen to be similar, means that botanic key construction is typically much more difficult than key construction from other (e.g. industrial) data. Often measurements do not form the clearly separable, compact clusters beloved of key constructors. These overlapping clusters are termed "poorly separated" clusters. It would be preferable if at least some of the data used to compare the methodologies was of this type.

It is also important that the characteristics be statistically independent. This is discussed further in Appendix E.

It would also be useful, (but not necessary), if the data sets had been previously used in other methodologies, in that a wider range of comparisons may then be available.

### 5.2.2 Data Requirements for the accurate identification of Botanic Specimens

Kidd comments that, if the key is to be of use to a user:

---

<sup>1</sup>Sneath, P. H. A., 'Philosophy and method in biological classification', in Felsenstein, J. (Ed)., *Numerical Taxonomy*, Springer-Verlag, Berlin, 1983, p. 27. For a brief discussion on cladistic taxonomy in relation to inductively derived keys, see section 5.1.2 of this thesis.

... it is vital that ... there is a high degree of cognitive compatibility between user and system. It must employ similar knowledge structures.<sup>1</sup>

If the key is to be accurate, the data from which it is constructed should have been collected in a statistically acceptable manner, such that each species or taxa is adequately represented, and no class of specimens observable in the field is under-represented.

The data should contain no data entry or mis-classification errors.

### 5.3 An examination of Data sets for use in Methodology Comparison

The suitability of some data sets for use in methodology comparison, given the requirements of section 5.2.1, are examined in section 5.3.1.

The suitability of these data sets for use in species identification (in the light of the requirements listed in section 5.2.2) is examined in section 5.3.2.

#### 5.3.1 Choice of data for suitability for use in methodology comparisons.

Of the sets of data examined for use in these series of comparisons, two sets of data were selected as most nearly meeting the requirements. These were the *Acaena* and *Danthonia* data.<sup>2</sup>

Both sets of data met the requirement that they were measurements of actual specimens gathered from botanic sources, and did not contain manufactured data.<sup>3</sup>

---

<sup>1</sup>Kidd, A. L., 'Human factors in expert systems', in Coombes, K., (Ed.), *Proceedings of the Ergonomic Society Conference 1983*, Taylor and Francis, London, 1983, (not seen), quoted by Gammack, J. G., 'Modelling expert knowledge using cognitively compatible structures', in *Third International Expert Systems Conference*, Learned Information (Europe) Ltd, London, 1987, p. 192.

<sup>2</sup>The origins of these sets of data are noted in section E.1 of Appendix E of this thesis.

<sup>3</sup>For more detail, see section E.2 of Appendix E of this thesis.

Both sets of data contained both parametric and non-parametric sets of characteristic observations.<sup>1</sup>

Both sets of data contained outliers.<sup>2</sup>

In both sets of data the characteristics observed were, in most cases, reasonably statistically independent.<sup>3</sup>

The requirement that some data observations be incomplete was met more than adequately by the *Acaena* data, in which approximately three-quarters of the specimens did not have complete data. The *Danthonia* data set was complete.

Both the *Acaena* and *Danthonia* data contained some cases where the number of specimens per species was below the amount statistically preferred, this occurring more frequently in the *Acaena* data than in the *Danthonia* data.<sup>4</sup>

Both sets of data contained portions of the data which were statistically poorly separated.<sup>5</sup>

Both data sets have been previously used by other authorities.<sup>6</sup>

### 5.3.2 An examination of data sets for use in accurate botanic specimen identification

Both sets of data were judged to have met the requirement that there be a high degree of cognitive compatibility between the users and the system, as in the case of both sets of data the characteristics to be observed were specified by an acknowledged expert in the field.<sup>7</sup>

---

<sup>1</sup>This is discussed in greater detail in section E.4.1 of Appendix E of this thesis.

<sup>2</sup>The presence of outliers is discussed in greater detail in section E.4.2 of Appendix E of this thesis. Possible data entry errors are discussed in section E.4.2.1 of Appendix E. The possibility of anomalous specimens or anomalous classifications of specimens is discussed in section E.4.2.2 of Appendix E.

<sup>3</sup>For more information, see Section E.3 of Appendix E of this thesis. In particular, Tables 69 and 70 of this Appendix list correlations between the observed characteristics.

<sup>4</sup>E.g. for an extreme example, see the previous discussion in section 3.1.3.2.1 (including Table 5) of this thesis.

<sup>5</sup>For more specific information on this, see sections A.1.2 and A.2.1 of Appendix A of this thesis, including Figures 32 to 34. However most of the results obtained in Appendix A would suggest the existence of poorly separated data. See also an extreme (artificial) example of this type of data in Table 1 of this thesis.

<sup>6</sup>See last paragraph of section E.2 of this thesis.

<sup>7</sup>See section E.1 of Appendix E of this thesis.



Whilst the condition that the data collection process should be statistically acceptable was vital for the production of a useful and accurate identification key, it could not be verified in the cases of the *Acaena* and *Danthonia* data (as is very often the case for botanic data collections). It could not be verified in these cases of the because the circumstances of the collection of the data were not known to this author. However even if the circumstances were known, it is a fact that the known geographical distribution of both *Acaena* and *Danthonia* make a statistically acceptable sampling process financially difficult.<sup>1</sup> This difficulty is often balanced in practice by the collecting expert using his or her knowledge to ensure that the collection is the most representative sample that it is financially possible to collect.<sup>2</sup> It was felt that, despite this uncertainty, the considerable knowledge of the experts supervising the collection of the data made the likelihood of adequate collection high.

Regarding data entry or mis-classification errors, in both data there were a small number of possibly suspect readings and specimens. Since neither the original data nor the specimens from which the data was obtained were available to this author, this uncertainty could not be resolved.<sup>3</sup>

The vast majority of the data, however, fitted these requirements.

---

<sup>1</sup>A map showing the world-wide distribution of the genus *Acaena* is to be found in: Humphries, Christopher J. and Parenti, Lynne R., *Cladistic Biogeography*, Clarendon Press, Oxford, 1989, Figure 1.5, p. 6.

<sup>2</sup>It should be noted that the chance of discovering a new species or taxa is less when present knowledge is used to collect a restricted range of specimens for analysis, compared with the chance of discovery of novelty if a statistically acceptable sampling is performed. However the practical difficulties in obtaining a truly representative botanic data sample are demonstrated in Evans *et al.*'s revision of their key for the British sub-montane plant communities. In this case the revision included 631 limestone samples to supplement the data from which the original key was obtained, it having been noticed that (because of an objective sampling method) the original 871 previous samples had been obtained from a predominantly acidic sub-strata. Even then they note (p. 2) that the key lacks data from mountain tops, cliffs, streams, lakes and flushed areas, plus a lack of data from complete regions such as the area north of the Central Lowlands in Scotland, and the North York Moors in England. This is an illustration that is difficult to impossible in practice to obtain a statistically acceptable sample of data of a complete area in botany due to restrictions such as finance, personnel and time. See: Evans, D.F., Hill, M.O. & Ward, S.D., *A dichotomous key to British submontane vegetation*, Occasional Paper No. 1, Institute for Terrestrial Ecology, Bangor, North Wales, 1977.

<sup>3</sup>For further discussion of possible anomalies see section E.4.2 of Appendix E.

## 5.4 Training and Test Sets of Data

Many of the methodologies which were to be compared with the methodology suggested in this thesis used one set of data for training, and another set of data for testing the accuracy of the training achieved by use of the first set of data. These sets were obtained by dividing each of the *Acaena* and *Danthonia* data into two sets, one to be used for training, one to be used for testing purposes. Usually the selection was arranged so that 80% of the specimens from the original data sets were allocated to the training set, and 20% to the test set.<sup>1</sup>

The method of dividing the data posed some methodological problems. Statistical considerations suggested that the methodology known as *proportional stratified sampling* would be the best method to be used when extracting the test data set from the complete data set, (leaving the remainder to be used as the training data set).<sup>2</sup> However:-

it has been algebraically demonstrated by mathematicians that the estimate of a population mean based on stratified sampling has the greatest precision (smallest standard error) when the sample sizes for the strata are in the same ratio as the products of the standard deviation of the stratum and stratum size.<sup>3</sup>

This posed a problem in the case of the botanical data, as there is not one characteristic per stratum to be sampled, but 41 in the case of the *Danthonia* data, and 31 in the case of the *Acaena* data. Also in the case of the *Acaena* data, some of the characteristics of some of the species have no useful data at all, and thus a standard deviation for this characteristic/species combination could not be calculated.

---

<sup>1</sup>If this ratio of split was not used in any particular methodology, it is documented in the sections associated with the results which were obtained by the use of that methodology.

<sup>2</sup>Edgington discusses the problem of sampling from a *finite* data set (as opposed to the more usually considered (but impractical) *infinite* data set) in chapter three of his book; see: Edgington, Eugene S., *Statistical Inference: The Distribution-free Approach*, McGraw-Hill Book Company, New York, 1969, pps. 21-47. A rather more mathematical treatment of stratified sampling is given in Steele, Robert G. and Torrie, James H., *Principles and Procedures of Statistics A Biometrical Approach*, Second Edition, McGraw-Hill International Editions, Singapore, 1987, pps. 560-565.

<sup>3</sup>Edgington, p. 31.

Added to these requirements was the preference for the training and test data sets being able to be 'reconstructed' from the original data set at any time, preferably with the same 'reconstruction' being able to be obtained on different computing platforms.<sup>1</sup>

These considerations led to the chosen treatment, where each group of measurements from a specimen were allocated to the data file together, but the specimens were allocated in a random order, the specimens of each group being kept together in a block. This distribution of data was retained, and became the reference data for subsequent runs.

The data were then split by allowing a computer program to allocate every tenth set of characteristic measurements/classifications (one specimen) to either the test or training set. The computer user was given the choice of which tenths were allocated to which data sets. This produced a split between the test and training sets which was reproducible, and perhaps somewhat similar in spirit to a proportional stratified sample.<sup>2</sup>

However even this system was found to produce some problems on occasions. Sometimes an entire species was omitted from one of the training or test data sets. Understandably, this occurred particularly often when there were not many specimens representing a particular species. This omission caused some of the comparison methodologies (as implemented) to fail. In particular, whilst some of the comparison methodologies could handle a situation where there was no useful data for a characteristic/species combination, no learning/test run was possible when (e.g.) SAS was presented with data in which a species was missing from either the training or test files. This was generally a problem with the way the comparison methodology had been implemented, rather than a limitation inherent in the methodology itself. To prevent unnecessarily disadvantaging these methodologies, the split data sets were adjusted so that each species was represented in each

---

<sup>1</sup>This effectively ruled out the often-used stratagem of a computerised 'random number' generator which starts from a fixed and reproducible starting point, generating the same series of 'random numbers' each time.

<sup>2</sup>These were written in Pascal, using the transportable Pascal system developed as part of this project; see section 4.7 b) of this thesis.

of the split sets of data. This was done by randomly selecting a specimen of the deficient species from the other data set, and adding it to the deficient data set.<sup>1</sup>

To help ensure that any randomly chosen sets did not, by chance, favour a particular methodology, this process was repeated a number of times to form multiple learning/test data sets from the original *Acaena* and *Danthonia* data sets.<sup>2</sup> These data sets were presented to the methodology being tested, and an average of the identification rates so obtained were then taken as an indication of the accuracy of the particular methodology under test.

For reference purposes, this methodology for obtaining the test and training data from the original data set will be referred to in the rest of this thesis as an 80/20 approximate stratified data split.

## 5.5 Summary

Three types of dendrograms are commonly used in botanic work; genealogical, cladistic, and identification dendrograms. Of these three, the methodology outlined in this thesis was postulated to assist best the construction of identification dendrograms. Identification dendrograms will be referred to as *keys* in the remainder of this thesis, to help distinguish them from the other types of dendrograms.

Various requirements for the botanic data to be used when comparing alternative methodologies for the identification of botanic specimens have been suggested.

After examination of the *Acaena* and *Danthonia* data sets it was concluded that, considering the above requirements, these data sets would seem to be reasonably representative of the type of problems inherent in many sets of botanic data intended for classificatory purposes, and thus would prove suitable data sets

---

<sup>1</sup>This had the effect of adjusting the total number of specimens in the training or test data sets up or down by one. The total number of specimens in the various training and data sets thus varies slightly between runs.

<sup>2</sup>Usually 8 data sets were used in the case of the *Danthonia* data, 7 in the case of the *Acaena* data.

**Inductive Categorisation, Dendrograms and Botanical Data**

**for comparing the classificatory methodologies with systems intended for use in developing keys for use with botanic data.**

Given the limitations usually inherent in the collection of these types of data, it was also concluded that these data sets were reasonably suitable for the production of accurate and useful identification keys.

# INDUCTIVE CATEGORISATION

## KEY CONSTRUCTION

This chapter discusses the results obtained from test runs of the Selecta-key system. The results are discussed and compared with alternative classification methodologies.

In section 6.1 of the following chapter comments are made on the test runs of Selecta-key and the results compared with the results with those obtained by use of the commercial product *1st Class*.<sup>1</sup>

Section 6.2 comments on the test runs of the Selecta-key implementation and compares the results with those obtained by the use from comparative runs obtained with Collier's implementation of Quinlan's entropy-based methodology ID3 and Quinlan's C4.5<sup>2</sup>.

Section 6.3 notes the comparative times used by key construction methodologies ID3 and Selecta-key.

The issue of when to use randomisation tests in place of the parametric normal assumption in the case of the *Acaena* and *Danthonia* data is examined in section 6.4.

An examination of the Selecta-key results against the background of runs of the *Acaena* and *Danthonia* data with discriminant analysis<sup>3</sup>, various clustering procedures (which in some cases were followed by a canonical discriminant analysis)<sup>4</sup>,

---

<sup>1</sup>The data used in the test runs is mainly Orchard's *Acaena* and Collier's *Danthonia* data, which are examined in some detail in chapter 5 and Appendix E of this thesis. The Selecta-key key construction methodology is compared to Quinlan's ID3 key construction methodology, against the background of some statistical methodologies (discriminant analysis (Appendix D) and several clustering methodologies (Appendix A)), two neural net methodologies (Appendix B), and a simplification of the Selecta-key approach referred to as the voting methodology (Appendix C).

<sup>2</sup>Collier's ID3 implementation *TL* became available subsequent to the earlier work employing *1st Class*; it was preferred to *1st Class*, as it was a more versatile implementation.

<sup>3</sup>SAS implementations of discriminant analyses were used; see Appendix D.

<sup>4</sup>Several SAS implementations of different clustering algorithms were used for comparison; see Appendix A for further discussion of the methodologies employed and the results obtained.

two implementations of neural net methodology<sup>1</sup>, and a simpler variation of the Selecta-key methodology referred to as the voting methodology<sup>2</sup> can be found in section 6.5.

The overall results obtained are compared and discussed in section 6.6.

Section 6.7 presents a summary.

## 6.1 Comparison of Results obtained from Selecta-key and 1<sup>st</sup> Class.

It can be convenient to construct "best" and "alternate" keys. The "best" key would use all the information available in the training data set. An alternate key may be convenient for use in cases where seasonally available characteristics are not apparent. Section 6.1.1 examines the "best" key option. Section 6.1.2 examines attempts to produce alternate keys.

### 6.1.1 Key construction — Selecta-key and 1<sup>st</sup> Class.

The application of probabilistic methods via the first prototype of the Selecta-key system was compared initially with the results obtained by Collier, who used the commercially available 1<sup>st</sup> Class package, which included an inductive classification algorithm.<sup>3</sup>

The data initially examined consisted of measurements of specimens of the *Acaena ovina* complex.

When the *Acaena* data was examined using the small-sample parametric approach via Selecta-key's first prototype, a problem similar to that represented in Table 3 was encountered in the monothetic single access botanical key produced. The key produced is shown in Figure 21.

---

<sup>1</sup>See Appendix B for a discussion of some neural net theory together with the results of the methodologies employed.

<sup>2</sup>See Appendix C for a discussion of the voting methodology, and the results obtained by the application of this methodology.

<sup>3</sup>The inductive classification algorithm was believed to be based on Quinlan's ID3 algorithm.

1	Fruit +/- glabrous	
2	Fruit spines +/- equal in length	
3	Stamens up to 2.8 mm long	<i>Acaena agnipila</i> var. <i>aequispina</i>
3*		<i>Acaena agnipila</i> var. <i>protenta</i>
2*		
4	Scape hair density	{ <i>Acaena echinata</i> var. <i>robusta</i>
4*		{ <i>Acaena ovina</i> var. <i>ovina</i>
		<i>Acaena echinata</i> var. <i>echinata</i>
1*		
5	Hairs on bottom of leaflet +/- confined to midrib & veins	
6	Length of short spines <1.6 mm	<i>Acaena agnipila</i> var. <i>agnipila</i>
6*		<i>Acaena agnipila</i> var. <i>tenuispica</i>
5*		
7	Hairiness of leaflet on top	
8	Fruit with ridges	<i>Acaena echinata</i> var. <i>retrorsumpilosa</i>
8*		<i>Acaena echinata</i> var. <i>subglabricalyx</i>
7*		
9	Up to 5 stamens	<i>Acaena ovina</i> var. <i>velutina</i>
9*		<i>Acaena echinata</i> var. <i>tylacantha</i>

Figure 21 — *Acaena* Key produced by Selecta-key system.

Note that it has not been necessary to "prune" this key. Note also that the key is generally similar to Figure 6 from Collier's paper.<sup>1</sup> (Collier's Figure 6 is shown below, by permission, as Figure 22.)

1.	Fruit ± glabrous	
2.	Fruit spines ± equal in length	
3.	Up to 5 stamens	<i>Acaena agnipila</i> var. <i>aequispina</i>
3*.		<i>Acaena agnipila</i> var. <i>protenta</i>
2*.		
4.	Hairs on bottom of leaflet ± confined to the midrib and veins	
5.	Stamens up to 3mm long	<i>Acaena echinata</i> var. <i>echinata</i>
5*.		<i>Acaena echinata</i> var. <i>robusta</i>
4*.		<i>Acaena ovina</i> var. <i>ovina</i>
1*.		
6.	Fruit spines ± equal in length	
7.	Inflorescence branched	<i>Acaena agnipila</i> var. <i>agnipila</i>
7*.		<i>Acaena agnipila</i> var. <i>tenuispica</i>
6*.		
8.	Fruit with ridges	<i>Acaena echinata</i> var. <i>retrorsumpilosa</i>
8*.		
9.	Up to 5 stamens	
10.	Hairs on petiole ± appressed	<i>Acaena ovina</i> var. <i>velutina</i>
10*.		<i>Acaena echinata</i> var. <i>subglabricalyx</i>
9*.		<i>Acaena echinata</i> var. <i>tylacantha</i>

Figure 22. The summary decision key from Collier's Figure 6.

While this key is generally similar to Figure 21, there are some differences, and it is instructive to examine them.

Firstly, the data does not allow a distinction to be made between *Acaena echinata* var. *robusta* and *Acaena ovina* var. *ovina* at the 5% level, the situation being similar to that postulated in

<sup>1</sup> Collier, *Inductive Inference for Botanical Keys*, p. 135



Figure 8. This is partly explained by there being data for only two examples of *Acaena echinata* var. *robusta*, and in six of the thirty factors, data on one or both of these specimens was missing. Also, it is mentioned by Collier that Orchard notes that *Acaena ovina* var. *ovina* is possibly of hybrid origin, and so may be expected to be variable.<sup>1</sup> A combination of a small number of samples in one group, and considerable variability in the second, have combined to make the null hypothesis (that these two are drawn from the same sample) impossible to reject at the 5% level. With more examples, a distinction may be possible. However, considering the present data alone, it is suggested that the distinction drawn by 1<sup>st</sup> Class between these two taxa can not be supported statistically, considering the available evidence.

For similar reasons, it is suggested that question 9 in Figure 21 is preferable to question 10 in Figure 22. The former can be separated at the 1% level, the latter pair cannot be separated statistically at the 5% level.

Secondly, in separating *Acaena agnipila* var. *aequispina* from *Acaena agnipila* var. *protenta*, 1<sup>st</sup> Class chooses the factor "up to 5 stamens", and the small-sample parametric approach method chooses "stamens up to 2.8 mm long". The t test results for these two factors are shown in Tables 10 and 11.

<b>Factor</b>	Number of Stamens
<b>Taxa</b>	<i>Acaena agnipila</i> var. <i>aequispina</i> <i>Acaena agnipila</i> var. <i>protenta</i>
<b>t test between means</b>	5.05
<b>Significance level</b>	<1%

Table 10 — t test results

<sup>1</sup>Collier, *Inductive Inference for Botanical Keys*, pps. 135 - 137.

<b>Factor</b>	Length of Stamens
<b>Taxa</b>	<i>Acaena agnipila</i> var. <i>aequispina</i> <i>Acaena agnipila</i> var. <i>protenta</i>
<b>t test between means</b>	5.92
<b>Significance level</b>	<1%

Table 11 — t test results

It will be noted that whilst the "length of stamens" factor was chosen by the t test method because it was possible to reject the null hypothesis at a higher level of significance, either choice exhibits a satisfactory significance, and either are reasonable as a question in the botanical key. However the information and certainty that either is a reasonable choice is only available to the expert via an approach such as that used by Selecta-key, where the results are fed directly back to the expert. If *1<sup>st</sup> Class* is used, the researcher is obliged to rely on other (probably circumstantial) evidence to assist in "pruning" decisions.

The differences between the choices made by *1<sup>st</sup> Class* in selecting questions 4, 6, 7 and 8 of Figure 22, compared with the choice of questions 4, 5, 6 and 7 of Figure 21 are for similar reasons; in each case the statistical method chose higher "t test" values, but either choice would be acceptable.

### 6.1.2 Alternate Key construction — Selecta-key and *1<sup>st</sup> Class*.

One advantage of any computerised approach is the ease of constructing alternate keys.

Suppose that no information is available on *Acaena* flowers, but fruit data is available. The statistical method produces a key as shown in Figure 23. Again no "pruning" is necessary. Note that flower data is necessary to separate *Acaena echinata* var. *tylacantha* and *Acaena echinata* var. *velutina* in a statistically significant manner.

1	Fruit +/- glabrous	
2	Fruit spines +/- equal in length	
3	Hairiness of leaflets on top	<i>Acaena agnipila</i> var. <i>aequispina</i>
3*		<i>Acaena agnipila</i> var. <i>protenta</i>
2*		
4	Scape hair density	{ <i>Acaena echinata</i> var. <i>robusta</i>
		{ <i>Acaena ovina</i> var. <i>ovina</i>
4*		<i>Acaena echinata</i> var. <i>echinata</i>
1*		
5	Hairs on bottom of leaflet +/- confined to midrib & veins	
6	Hairiness of leaflets on top	
7	Fruit with ridges	<i>Acaena echinata</i> var. <i>retrorsumpilosa</i>
7*		<i>Acaena echinata</i> var. <i>subglabricalyx</i>
6*		{ <i>Acaena echinata</i> var. <i>tylacantha</i>
		{ <i>Acaena ovina</i> var. <i>velutina</i>
5*		
8	Length of short spines <1.6 mm	<i>Acaena agnipila</i> var. <i>agnipila</i>
8*		<i>Acaena agnipila</i> var. <i>tenuispica</i>

Figure 23 — *Acaena* Classification Key, constructed without Flower Data.

A similar attempt to construct a key for *Acaena* when no fruit data is available, was less successful. Incomplete separation occurred frequently.<sup>1</sup>

The "no fruit" key only allowed *Acaena agnipila* var. *tenuispica* to be uniquely separated. The following; *Acaena agnipila* var. *agnipila*, *Acaena agnipila* var. *protenta*, *Acaena echinata* var. *tylacantha*, *Acaena agnipila* var. *aequispina*, and *Acaena ovina* var. *ovina* could be separated sometimes, and sometimes not.

*Acaena echinata* var. *robusta*, *Acaena echinata* var. *velutina*, *Acaena echinata* var. *retrorsumpilosa*, *Acaena echinata* var. *subglabricalyx* and *Acaena echinata* var. *echinata* could never be separated.

With no fruit and no flower data available, the statistical method could not identify any taxa uniquely. The smallest group consisted of five taxa. This was because the data was statistically inadequate to produce a botanical key separating all taxa in the absence of flower and fruit data. This was an example of the type of situation noted by Pankhurst when he comments:

<sup>1</sup> A diagrammatic example of the type of situation that led to this lack of separation is shown in Figure 19 of this thesis. In this case, using an  $\chi^2_{split} = S1$ ,  $\alpha$  and  $\delta$  are statistically distinct, but  $\beta$  may not be separated from either of them. Hence  $\beta$  will appear on both branches of the botanical key following this decision. This occurred frequently in the attempt noted above.

there may be taxa which are not distinguishable with available characters, and then a *partial key* can be made up. The ultimate leads of such a key may give the names of more than one taxon, instead of one only. When such a key is in use, one may find that it is only possible to reach a short list of alternative identifications instead of a definitive one.<sup>1</sup>

The use of Selecta-key allowed this type of *partial key* to be produced. By contrast, with the same data, 1<sup>st</sup> Class produced the key shown in Figure 24, following:

```

----- start of rule -----
1: leaflet_hb??
2:   midrib/vein:leaflet_no??
3:     <16.75:leaflet_ht??
4:       <1.25:petiole_ho??
5:         <1.50:leaflet_w??
6:           <8.25:leaflet_ht??
7:             <0.50:-----ech_ech
8:             >0.50:-----ech tyl
9:             >8.25:-----ech tyl
10:        >1.50:petiole_ho??
11:          <2.50:leaflet_l??
12:            <10.75:-----ech_sub
13:            >10.75:-----ovi_vel
14:          >2.50:serrations??
15:            <10.25:leaflet_w??
16:              <6.25:leaflet_no??
17:                <12.50:-----ech_ret
18:                >12.50:leaflet_w??
19:                  <5.25:-----ech_ech
20:                  >5.25:leaflet_l??
21:                    <9.25:-----ech_ech
22:                    >9.25:-----ech_ret
23:                      >6.25:leaflet_l??
24:                        <10.50:-----ech_ech
25:                        >10.50:-----ech_sub
26:                      >10.25:leaflet_l??
27:                        <8.50:-----ech_sub
28:                        >8.50:leaflet_w??
29:                          <8.00:-----ech_ret
30:                          >8.00:leaflet_w??
31:                            <10.25:-----ech_sub
32:                            >10.25:-----ech_ret
33:                      >1.25:serrations??
34:                        <9.50:-----ech tyl
35:                        >9.50:-----ovi_vel
36:                    >16.75:leaflet_ht??
37:                      <0.25:leaflet_w??
38:                        <8.00:-----ech_rob
39:                        >8.00:-----ech_ech
40:                      >0.25:depth_serr??
41:                        <0.54:-----ech_ech
42:                        >0.54:-----ovi_ovi
43:    all_over:leaflet_ht??

```

<sup>1</sup>Pankhurst, Richard J., *Practical taxonomic computing*, Cambridge University Press, Cambridge, 1991, p. 95. The italics were in the original text.

```

44:      <1.50:stipule_w??
45:      <1.25:depth_serr??
46:      <0.63:leaflet_l??
47:      <8.00:leaflet_w??
48:      <4.50:-----ovi_ovi
49:      >4.50:-----ovi_vel
50:      >8.00:depth_serr??
51:      <0.54:-----agn_ten
52:      >0.54:-----agn_aeq
53:      >0.63:-----ovi_ovi
54:    >1.25:petiole_ho??
55:      <1.50:leaflet_w??
56:      <8.00:-----agn_ten
57:      >8.00:leaflet_w??
58:      <8.75:-----agn_pro
59:      >8.75:-----agn_ten
60:      >1.50:leaflet_w??
61:      <6.00:-----agn_agn
62:      >6.00:leaflet_w??
63:      <7.25:-----ech tyl
64:      >7.25:-----ovi_vel
65:    >1.50:serrations??
66:      <11.25:leaflet_no??
67:      <18.50:leaflet_no??
68:      <15.50:-----ech tyl
69:      >15.50:-----agn_aeq
70:      >18.50:leaflet_w??
71:      <6.25:-----ovi_vel
72:      >6.25:-----agn_pro
73:    >11.25:leaf_len??
74:      <11.75:leaflet_w??
75:      <8.25:leaflet_no??
76:      <20.75:leaflet_l??
77:      <11.50:depth_serr??
78:      <0.46:-----agn_aeq
79:      >0.46:-----agn_agn
80:      >11.50:-----agn_aeq
81:      >20.75:-----agn_agn
82:      >8.25:-----ovi_vel
83:    >11.75:stipule_l??
84:      <3.75:leaflet_w??
85:      <8.75:-----agn_ten
86:      >8.75:-----agn_aeq
87:      >3.75:leaflet_no??
88:      <14.75:-----agn_ten
89:      >14.75:-----agn_agn
    ---- end of rule ----

```

Figure 24 — Key from 1<sup>st</sup> Class, using no-flower, no-fruit *Acaena ovina* data.

It may be noted that a key 89 lines long has been produced from this data set of 81 specimens. A pruning algorithm could be used to try and improve the key, but even this may be of doubtful efficacy, as there is an average of less than 2 specimens per leaf node, and some taxa may be represented by only single example leaf nodes. The problems produced by this data may be regarded as similar to the problems produced by noisy data.

It is suggested that the ability to produce a seemingly adequate key from inadequate data is not a desirable property in a key producing program.

Another problem found with *1<sup>st</sup> Class* was an inability to produce even a partial key if two taxa of different species in the data had an identical set of characteristics. By contrast, under these circumstances Selecta-key still produced a botanically useful key.

## 6.2 Comparing Selecta-key's *Acaena* and *Danthonia* Keys with existing keys.

"Best" keys for both the *Acaena ovina* complex and the Tasmanian *Danthonia* genus were prepared using the Selecta-key process.<sup>1</sup>

The *Acaena* and *Danthonia* data were split into training and test data sets, using the 80/20 approximate stratified split methodology.<sup>2</sup>

The keys were constructed using the training sets of data.

The resultant keys were then compared with other keys constructed for the same species by both other computer methods and human experts, using the test sets of data.

Each selected specimen was submitted for identification to the key under consideration. If the specimen was unable to be identified (the data is realistic in that many specimens have only partial data available), the specimen was categorised as "unclassifiable".

The results of these comparisons are shown below. Results are given for the *Acaena ovina* complex first, and for the *Danthonia* genus next.

---

<sup>1</sup> See section 6.1 for a discussion of "best" and "alternate" keys.

<sup>2</sup> For more detail, see section 5.4 of this thesis.

### 6.2.1 Results obtained using Selecta-key with the *Acaena* data

In this section the identification rates obtained with the *Acaena* data by the application of Collier's key are given in section 6.2.1.1. Quinlan's C4.5 (section 6.2.1.2) and Orchard's key (section 6.2.1.3) are then applied to the same data. An attempt was made to duplicate Orchard's key using the Selecta-key system (section 6.2.1.4). The result of the Selecta-key process is given in section 6.2.1.5.

#### 6.2.1.1 *Acaena* Data — Collier's Summary Key

The first key examined was the summary key presented by Collier, Figure 22. This key allows only one characteristic to be used per split, and each identification to occur only once in the key. These restrictions ensure a straightforward, simple, easy to use key, but as may be expected the restrictions also help ensure a lower ability to correctly identify individual specimens.<sup>1</sup> However the rate of correct identification was still good, as may be seen from Table 12.

Correctly Classified	Incorrectly Classified	Unable to Classify
63%	15%	22%

Table 12 — Classification rate obtained by use of Figure 22 key.

#### 6.2.1.2 *Acaena* Data — Quinlan's C4.5 Algorithm

Next a key produced by Quinlan's C4.5 algorithm was used.<sup>2</sup> This key was first presented by Collier, and is reproduced here, with permission, as Figure 25.

<sup>1</sup>Matters relating to this are discussed in section 3.2.2.2 of this thesis.

<sup>2</sup>For a comparison of results of Quinlan's C4 and another very interesting learning approach (genetic algorithms) not otherwise considered in this thesis, see McCallum, R. Andrew and Spackman, Kent A., 'Using Genetic Algorithms to Learn Disjunctive Rules from Examples', in Porter, Bruce and Mooney, Raymond, *Machine Learning: Proceedings of the Seventh International Conference*, Morgan Kaufmann, San Mateo, 1990, pps. 149-152; also Bonelli, Pierre, Parodi, Alexandre, Sen, Sandip and Wilson, Stewart, 'NEWBOOLE: A Fast GBML System', in Porter et. al., pps. 153-159.

This key also uses only one characteristic per split, but allows multiple appearances of the same species in the key. This contributes towards the very good identification rate obtained, see Table 13.

- 1. Fruit ± glabrous
  - 2. Fruit spines ± equal in length
    - 3. Up to 5 stamens
      - 3\*. *Acaena agnipila* var *aequispina*  
*Acaena agnipila* var *protenta*
    - 2\*.
      - 4. More than 5 stamens
        - 4\*. *Acaena echinata* var *robusta*
  - 5. Hairs on base of leaflets ± confined to veins/midrib
    - 5\*. *Acaena echinata* var *echinata*
  - 5\*. Hairs on base of leaflets ± evenly distributed all over
    - 5\*. *Acaena ovina* var *ovina*
- 1\*.
  - 6. Fruit with ridges
    - 6\*. *Acaena echinata* var *retrorsumpilosa*
  - 7. Fruit spines ± equal in length
    - 8. Inflorescence branched
      - 8\*. *Acaena agnipila* var *agnipila*  
*Acaena agnipila* var *tenuispica*
    - 7\*.
      - 9. More than 5 stamens
        - 9\*. *Acaena echinata* var *tylacantha*
      - 10. Inflorescence branched
        - 10\*. *Acaena agnipila* var *agnipila*
      - 11. Hairs on petiole ± appressed
        - 11\*. *Acaena ovina* var *velutina*
      - 12. Up to 12 leaflet serrations
        - 12\*. *Acaena echinata* var *subglabricalyx*  
*Acaena ovina* var *velutina*

Figure 25. The decision key produced by the C4.5 algorithm.

Correctly Classified	Incorrectly Classified	Unable to Classify
69%	5%	26%

Table 13 — Classification rate obtained by use of C4.5 key.

6.2.1.3 *Acaena* Data — Orchard's Key

Next the specimens were submitted to a key produced by Orchard's revision of the complex. The key differs from those considered so far in that multiple characteristics are used per split. Also there are examples of what may, in computer terms, be called IF...THEN characteristics, as opposed to the more usual IF...THEN...ELSE characteristics used in splits.



1. Flowers and fruits all arranged in a globular terminal head. Fruits 4 angled with 4 slender subequal spines, 1 at the apex of each angle. Stamens 2, cream. Creeping plants with long epigleal stolons.
  2. Calyx lobes persistent in fruit  $\pm$  fused at the base, spines long (1-2 cm) in fruit, fruiting head 2-3 cm diam. *A. anserinifolia* complex
  2. Calyx lobes deciduous in fruit, free at the base, spines short (1-3 mm) in fruit, fruiting head in 1 cm diam. *A. montana*
1. Flowers and fruits not in a head as above. Stamens 2-8, purple. Plants lacking (except occasionally in *A. X anserovina*) long stolons, forming tight clonal clumps.
  3. Fruits in globular heads with 3 or 4 fruits scattered on stem below,  $\pm$  4 angled or globular, 4-6 slender spines at the apex, and several smaller ones on the body of the fruit. 4. *A. X anserovina*
  3. Fruits in elongate interrupted cylindrical spikes, ovoid or with 3-4 longitudinal angles. Spines many, equal or unequal, scattered over the entire fruit. (*Acaena ovina* complex.)
  4. Leaflets densely and evenly appressed pilose on the under surface, moderately appressed pilose on the upper surface; fruit ovoid,  $\pm$  wrinkled, but in any case lacking longitudinal ridges formed by concrescence of thickened spine bases, spines slender.
    5. Spines of fruit unequal, 3-6 longer than the rest. 3. *A. ovina*
    6. Body of fruit and spines glabrous. var. *ovina*
    6. Body of fruit densely spreading pilose, spines  $\pm$  pilose at extreme base. var. *velutina*
    5. Spines of fruit  $\pm$  equal. 1. *A. agnipila*
    7. Fruit densely spreading pilose, spines glabrous, or pilose at extreme base only.
      8. Stamens (5-) 6 (-7). Length 2.5-3.5 mm, stipules 4.0-5.0 (-8.0) mm. spike  $\pm$  branched at the base. var. *agnipila*
      8. Stamens (3-) 4-5, length 1.5-2.0 mm, stipules 1.0-3.0 mm long, spike unattached. var. *tenuispica*
    7. Fruits and spines glabrous.
      9. Stamens 3-4 (-5), length 1.5-2.0 mm, stipules 2.0-3.5 mm long. var. *aequispina*
      9. Stamens (5-) 6 (-7), length 4.0 mm, stipules 4.0-5.0 long. var. *protenta*
  4. Leaflets with hairs confined to the major veins and/or midrib on the lower surface, glabrous or  $\pm$  sparsely pilose on the upper surface; fruit ovoid with all spines slender or with 3-4 longitudinal ridges formed by concrescence of the thickened bases of the 3-8 largest spines; spines always unequal. 2. *A. echinata* Nees
  10. Fruit and spines glabrous, largest spines with thickened bases.
    11. Stamens (2-) 4-5, length 1.5-2.0 mm, stipules 1.5-2.5 mm long. var. *echinata*
    11. Stamens 6-8, length 3.5-4.0 mm, stipules 3.0-5.0 long. var. *robusta*
  10. Fruit spreading pilose, larger spines with thickened bases or slender.
    12. Spines all slender, fruit ovoid without longitudinal ridges. var. *subglabricalyx*
    12. Spines (at least the longest ones) with thickened bases, fruit with 3-4 longitudinal ridges.
      13. Stamens (2-3-) 4-5, length 1.0-2.0 mm, stipules 2.0-3.0 mm long, spike unbranched. var. *retrorsumpilosa*
      13. Stamens (4-) 6, length 3.0-5.0 mm, stipules 4.0-5.0 mm long, spike usually branched at base. var. *tylacantha*

Figure 26 — Orchard's Key to the Australian Species and varieties of *Acaena*.<sup>1</sup>

These IF...THEN one-sided type of characteristic specifications are not available with ID3-type inductive algorithms. They are available only in unusual circumstances (and

<sup>1</sup>See Orchard, A. E., 'Revision of the *Acaena Ovina* A. Cunn. (Rosaceae) Complex in Australia', Trans. Roy. Soc. S. Aust. (1969), Vol. 93, pps. 91-109.

then only when supported by the data) with the Selecta-key approach. This additional versatility may contribute to the excellent rate of identification achieved by Orchard's key, see Figure 26. It will be noted that this key also identifies several *Acaena* taxa not included in the data available to this author, and so the key will be longer than the other *Acaena* keys presented in this thesis.

Correctly Classified	Incorrectly Classified	Unable to Classify
70%	28%	2%

Table 14 — Classification rate obtained by use of Orchard's key.

It will be noted that the use of multiple characteristics per split decreases enormously the number unable to be classified, but this very versatility also caused a higher rate of misclassifications in that all but one of the "difficult" specimens were classified.<sup>1</sup> By contrast, the two preceding keys "gave up", put the specimen into the "unclassifiable" category, and went on to classify (often correctly) an easier one, giving them a falsely higher rate of "correct" classifications if the "unclassifiable" category is omitted. The results obtained by use of Orchard's key are detailed in Table 14.<sup>2</sup>

#### 6.2.1.4 *Acaena* Data — Selecta-key Imitation of Orchard's Key

As an experiment, an attempt was made to duplicate Orchard's key using the Selecta-key programs. In this case, the recommended alternatives were ignored, and the characteristics at each decision level searched until one was found which most closely matched Orchard's chosen splitting characteristic. The characteristics were then re-examined to see if any other

<sup>1</sup>A "difficult" specimen is one where the characteristic(s) which most clearly separate the taxa are missing, and identification must be made using characteristics which show less separation, that is, overlap by a significant amount. Identification using these characteristics will inevitably lead to a higher error rate.

<sup>2</sup>It must be emphasised that the author does not regard himself as a botanical expert. An expert using Orchard's key may well have obtained a better result.

characteristics were available which produced the same taxa split, and hence could be used to reinforce the chosen decision.

When used to classify the test data, this key has a marginally higher rate of correctly classified specimens and a lower rate of incorrect classification than Orchard's key; although the difference between the two results are probably not statistically significant. However it achieves this by putting far more of the 'difficult' specimens in the 'unclassifiable' category, and classifying easier specimens. Overall, Orchard's key returns a better result. The results are shown in Table 15.

Correctly Classified	Incorrectly Classified	Unable to Classify
68%	24%	8%

Table 15 — Classification rate obtained by use of "imitation" key.

6.2.1.5 *Acaena* Data — Key derived from Selecta-key.

The last key examined for the *Acaena ovina* is a key produced with the aid of the Selecta-key process. This allows multiple characteristics to be used per split, and multiple occurrences of species in the key, if the expert using Selecta-key desires this state of affairs, and the data will support it. The resultant key is shown in Figure 27. The rate of success is shown in Table 16.

1. Fruit  $\pm$  glabrous.
  2. Leaflets with hairs confined to the major veins and/or midrib on the lower surface.
    3. Stamens up to 3.5 mm long. Up to 5 stamens. Up to 17 leaflets per leaf.  
Stipules up to 1.5 mm wide. *Acaena agnipila* var *echinata*
    3. Stamens more than 3.5 mm long. More than 5 stamens. More than 17 leaflets per leaf. Stipules more than 1.5 mm wide.  
*Acaena agnipila* var *robusta*
  2. Leaflets densely and evenly appressed pilose on the under surface.
    4. Less than 6 stamens. Stamens up to 2.95 mm long.
      5. Fruit spines  $\pm$  equal in length. Spines more than 1.42 mm long.  
Up to 27 spines. *Acaena echinata* var *aequispina*
      5. Some spines on the fruit markedly longer than others. Short spines up to 1.42 mm long. More than 27 short spines.  
*Acaena ovina* var *ovina*
    4. 6 or more stamens. Stamens more than 2.95 mm long.  
*Acaena echinata* var *protenta*
1. Fruit pilose.
  6. Fruit with 3-4 longitudinal ridges. *Acaena echinata* var *retrorsumpilosa*
  6. Fruit ovoid, lacking longitudinal ridges.
    7. Fruit spines  $\pm$  equal in length. If any long spines then less than 3.
      8. Spike unbranched. An average of up to 4.72 stamens.  
Stamen length up to 2.2 mm. Stipule width up to 1.6 mm.  
Stipule length up to 3.77 mm. *Acaena echinata* var *tenuispica*
      8. Spike branched. An average of more than 4.72 stamens.  
Stamen length more than 2.2 mm. Stipule width more than 1.6 mm. Stipule length more than 3.77 mm.  
*Acaena echinata* var *agnipila*
    7. 3 or more spines on the fruit markedly longer than the others.
      9. Stamens up to 2.55 mm long. An average of up to 4.76 stamens.  
Sepals up to 2.09 mm long.
        10. Leaflets glabrous or  $\pm$  sparsely pilose on the upper surface.  
Leaflets with hairs confined to the major veins and/or midrib on the lower surface. Hairs on the petiole spreading.  
Up to 15 leaflets. Scape  $\pm$  glabrous.  
*Acaena agnipila* var *subglabricalyx*
        10. Leaflets moderately appressed pilose on the upper surface.  
Leaflets densely and evenly appressed pilose on the lower surface. Hairs on the petiole appressed. More than 15 leaflets.  
Scape pilose. *Acaena ovina* var *velutina*
    9. Stamens more than 2.55 mm long. An average of more than 4.76 stamens.  
Sepals up to 2.09 mm long. *Acaena echinata* var *tylacantha*

Figure 27 — Selecta-key key for *Acaena ovina* taxa.

It will be noted that the use of the Selecta-key process supplied the expert with enough information to construct a key with a balance of accuracy and unclassifiability, permitting the highest overall rate of correct identification so far.

Correctly Classified	Incorrectly Classified	Unable to Classify
75%	13%	12%

Table 16 — Classification rate obtained by use of Selecta-key key.

This key was produced in about 20 minutes “wall time”.<sup>1</sup>

### 6.2.2 Results obtained using Selecta-key with the *Danthonia* data

The Selecta-key process had also been tested with data obtained by measuring all the *Danthonia* data available in the Tasmanian Herbarium. Less existing keys were available for comparison in the case of this data, but a comparison was made with a key first presented by Collier, see section 6.2.2.1. A further key for the same data produced by use of the Selecta-key methodology is examined in section 6.2.2.2.

#### 6.2.2.1 *Danthonia* data — Collier's Key

A key first presented by Collier is reproduced here, with permission, as Figure 28.<sup>2</sup> Taxa of the *Danthonia* genus are difficult to identify by use of a key, as much of the taxa are similar, and use of ID3-type algorithms often result in problems of the type found in Figure 24.

---

<sup>1</sup>Most of which was spent with the user reviewing alternatives presented by Selecta-key, whilst the computer waited for the user's selection. This time could reasonably be expected to vary markedly from user to user, depending on the user's familiarity with the subject being examined. Note that the key was produced in a very summarised format using abbreviated species and characteristic names. The 20 minutes did not include the time taken to decode the summarised results and then turn them into the neat format shown in Figure 27. This time could, of course, have been substantially reduced by altering the program so that it gave a decision tree automatically (as does ID3). This alteration is trivial as it would merely involve removing the question to the expert and automatically choosing the option which is mathematically (but possibly not practically) optimal. A less trivial, but still simple change would allow the automatic choice of matching characteristics for each splitting point. While these changes would improve the apparent 'efficiency' of the methodology (the timings reported in the Table 19 comparisons had to be made on a basis similar to this) it would eliminate a most important facet of the Selecta-key methodology, the ability of this system to combine the best of both worlds. Selecta-key combines the tireless mathematical ability of the computer with the common sense of the human expert to produce a result which would be quicker and practically superior to that which could have been easily produced by either functioning alone.

<sup>2</sup>P.A. Collier, 'Computer Key Generation from Quantitative Data', unpublished manuscript.

## Key Construction and Comparisons

- |    |  |    |                              |
|----|--|----|------------------------------|
| 1. | Lateral lobe of the lemma up to 5.2 mm long.   | 2  |                              |
|    | Lateral lobe of the lemma more than 5.2 mm long.   | 13 |                              |
| 2. | Up to 7 tufts of hairs in the upper row on the lemma.<br>Length of hairs up to 2.3 mm.   | 3  |                              |
|    | More than 7 tufts of hairs in the upper row on the lemma. Length of hairs more than 2.3 mm.  | 9  |                              |
| 3. | Awn up to 5 mm in length. If slightly more then the palea and body of lemma $\pm$ equal in length ( <i>Danthonia nitens</i> ) or upper row of lemma hairs more than 2 mm long ( <i>Danthonia pauciflora</i> ).                                 | 4  |                              |
|    | Awn more than 5 mm in length. If slightly less then the palea much longer than the body of lemma and upper row of lemma hairs up to 1.5 mm ( <i>Danthonia nudiflora</i> ).   | 6  |                              |
| 4. | Hairs in upper row on lemma up to 1.3 mm in length.<br>Up to 4 tufts of hairs in upper row. Up to 4 spikelets.   | 5  |                              |
|    | Hairs in upper row on lemma up to (1.5-) 2 mm or longer. 6 or more tufts of hairs in upper row.<br>(4-) 5 -10 (-14) spikelets.   |    | <i>Danthonia pauciflora</i>  |
| 5. | Length of body of the lemma less than 2.75 mm. Palea at least 1.2 times as long as the body of the lemma.<br>Awn length at least 1.45 times the body of the lemma.   |    | <i>Danthonia nivicola</i>    |
|    | Length of body of the lemma at least 2.75 mm. Palea less than 1.2 times as long as the body of the lemma.<br>Awn length less than 1.45 times the body of the lemma.  |    | <i>Danthonia nitens</i>      |
| 6. | 2 tufts of hair in the lower row on the lemma. Length up to 1 mm.  |    | <i>Danthonia nudiflora</i>   |
|    | 4 or more tufts of hair in the lower row on the lemma.<br>If less then some hairs more than 1 mm long.   | 7  |                              |
| 7. | Awn up to 2.25 times the length of the lateral lobe of the lemma. Lateral lobes 1.25-2.25 times the length of the body of the lemma. Upper row of hairs often placed in the upper third of the body of the lemma.                              | 8  |                              |
|    | Awn at least 2.5 times the length of the lateral lobe of the lemma. Lateral lobes often shorter, or occasionally slightly longer than the body of the lemma. Upper row of hairs often placed in the lower two thirds of the body of the lemma. |    | <i>Danthonia dimidiata</i>   |
| 8. | Number of florets 3-5 (-6-8). Awns exserted from the glumes for 0.25-0.4 (-0.5) of their length.   |    | <i>Danthonia penicillata</i> |
|    | Number of florets (4-) 6-10. Awns exserted from the glumes for (0.25-) 0.45-0.6 (-0.7) of their length.  |    | <i>Danthonia racemosa</i>    |

9. Panicle less than 50 mm in length. 10
- Panicle 50 mm or longer. If a little less then the marginal tuft of the ligule  $\pm$  absent (*Danthonia semiannularis*). 12
10. Glumes up to 6.75 mm long. If slightly more then hairs in the upper row of the lemma up to 3 mm long. *Danthonia pauciflora*
- Glumes more than 6.75 mm long. If slightly less then hairs in the upper row of the lemma more than 3 mm long. 11
11. 3 nerves extending more than half way up the glume. Callus hairs up to 1 mm long. Awn more than twice as long as the body of the lemma. Body of lemma up to 3.5 mm long. *Danthonia fortuneae-hibernae*
- 5-7-9 nerves extending more than half way up the glume. Callus hairs at least 1.4 mm long. Awn less than twice as long as the body of the lemma, frequently about the same length. Body of lemma at least 3.5 mm long. *Danthonia carphoides* var. *angustior*
12. Marginal tuft of the ligule often absent. If present then less than 10 hairs, the longest up to 1.4 mm. Ratio of column length to bristle length of the awn 0.3 or more. (4-) 5-7 florets per spikelet. *Danthonia semiannularis*
- Marginal tuft of the ligule well developed with (10-) 15-40 hairs, the longest more than 1.5 mm (up to 7 mm). Ratio of column length to bristle length of the awn up to 0.25. 2-3 (-4) florets per spikelet. *Danthonia gracilis*
13. Up to 6 (-7) tufts in the upper row of hairs on the lemma. If 6 (-7) then the awn up to 4 times the length of the body of the lemma. { *Danthonia pilosa*  
{ *Danthonia penicillata*  
{ *Danthonia racemosa*
- 8 or more tufts in the upper row of hairs on the lemma (or upper row  $\pm$  continuous hairs). If 6 then the awn more than 4 times the length of the body of the lemma (*Danthonia setacea*). 14
14. 2 tufts of hairs in the lower row on the lemma. *Danthonia laevis*
- 4 or more tufts of hairs in the lower row on the lemma. 15
15. Ratio of flat length to total length of lateral lobe of lemma up to 0.4. Ratio of length of awn to body of lemma 3.7 or more. 16
- Ratio of flat length to total length of lateral lobe of lemma at least 0.5. Ratio of length of awn to body of lemma up to 3.6. 17
16. Ratio of flat length to total length of lateral lobe of lemma up to 0.3. Glumes up to 2.5 mm wide. Length of body of lemma up to 3.2 (-3.7) mm. *Danthonia setacea*
- Ratio of flat length to total length of lateral lobe of lemma more than 0.3. Glumes more than 2.5 mm wide. Length of body of lemma more than 3.7 mm. *Danthonia caespitosa*

17.	Hairs not arranged in distinct tufts on lemma body. Distinct tufts of hair in two rows on the lemma body.	<i>Danthonia geniculata</i> 18
18.	Top row of hairs on the lemma in the lower three quarters of the lemma body. Panicle up to 45 mm long. Callus up to 0.4 mm long.	<i>Danthonia diemenica</i>
	Top row of hairs on the lemma in the upper quarter of the lemma body. Panicle more than 45 mm long. Callus more than 0.4 mm long.	19
19.	Length of the body of the lemma more than 5 mm. Panicle at least 70 mm long. Up to 4 florets.	<i>Danthonia procera</i>
	Length of the body of the lemma up to 5 mm. Panicle up to 70 mm long. More than 4 florets.	20
20.	Awn up to 13.5 mm in length. Panicle up to 70 mm long. Lemma body up to 4 mm long.	<i>Danthonia tenuior</i>
	Awn more than 13.5 mm in length. Panicle more than 70 mm long. Lemma body more than 4 mm long.	<i>Danthonia caespitosa</i>

Figure 28 — Collier's *Danthonia* key

The key produced by Collier uses multiple characteristics per split. It was obtained by much hand work, using repeated runs of Collier's implementation of Quinlan's ID3 methodology.<sup>1</sup> It used data which was in some cases restricted and manipulated to avoid problems of the type noted in Figure 24. It also allows identification of the same species of *Danthonia* at multiple places in the key. These factors made the key difficult to present in the same format as the rest of the keys shown here, hence a linear format is used for Figure 28.

The results obtained when test data was submitted to the key are summarised in Table 17.<sup>2</sup>

Correctly Classified	Incorrectly Classified	Unable to Classify
70%	12%	18%

Table 17 — Classification rate obtained by use of Collier's key.

<sup>1</sup>Results reported elsewhere suggest that ID3, by itself, produces results of the same order as back-propagation, (for back-propagation results with these data sets, see Tables 51 and 58 of this thesis); e.g. see Dietterich, Thomas G., Hild, Hermann and Bakiri, Ghulum, 'A Comparative study of ID3 and Backpropagation for English Text-to-Speech Mapping', in Porter, Bruce and Mooney, Raymond, *Machine Learning: Proceedings of the Seventh International Conference*, Morgan Kaufmann, San Mateo, 1990, pps. 24-31.

<sup>2</sup>The test set was 20% of the data, as discussed in section 5.4 of this thesis.



### 6.2.2.2 *Danthonia* data — Key derived from Selecta-key

The *Danthonia* data was also submitted to a key constructed by the expert using Selecta-key. The outline key was initially constructed in half an hour of "wall time",<sup>1</sup> during most of which time the computer was inactive whilst the expert considered and chose amongst the alternative characteristics available for each split, the 'response time' of the expert being much greater than that of the computer.<sup>2</sup> The user then repeated the exercise, filling in the multiple characteristics per split, where appropriate.<sup>3</sup> The abbreviated outline was then changed into the format shown in Figure 29. The overall time, (including typing the key into the computer) was several hours.

It will be noted that, in contrast to Figure 28, all *Danthonia* taxa are separable.

This key is presented with the split values suggested by Selecta-key, which currently prints to two decimal places. It will be noted that in some cases this accuracy may not be appropriate, however to present more truly the Selecta-key output, the original values have not been altered in this presentation of this key. The key is shown in Figure 29, and the results obtained from it in Table 18.

---

<sup>1</sup>Note that this time is within the maximum recommended time for continuous VDT work, see Helmut. T. Zwahlen, Andrea L. Hartmann, and Sudhakar L. Rangarajulu, 'Effects of rest breaks in continuous VDT work on visual and musculoskeletal comfort/discomfort and on performance,' in Gavriel Salvendy, (Ed.), *Human-Computer Interaction*, Elsevier Science Publishers, Amsterdam, 1984, p. 315.

<sup>2</sup>'Response time' here is taken to mean the time taken for the expert to assimilate the information supplied by the Selecta-key system, before providing a response. This is often neglected in computer systems, but is of significant importance, as noted by R. M. Balzer, 'Search for a Solution: A case study', in Donald E. Walker, and Lewis M. Norton, (Eds.), *Proceedings of the International Joint Conference on Artificial Intelligence*, Washington, 1969, p. 29.

<sup>3</sup>The initial key was in a single-characteristic, abbreviated format.

- |    |  |                              |
|----|--|------------------------------|
| 1. | Lateral lobe of the lemma up to 5.08 mm long.  | 2                            |
|    | Lateral lobe of the lemma more than 5.08 mm long.  | 14                           |
| 2. | Hairs in the upper row on the lemma up to 2.32 mm long.<br>Up to 7 tufts of hairs.   | 3                            |
|    | Hairs in the upper row on the lemma more than 2.32 mm long. More than 7 tufts of hairs.  | 9                            |
| 3. | Awn up to 5.11 mm long. Panicle up to 24.21 mm long. Glumes up to 7.51 mm long. Lateral lobes of the lemma up to 2.43 mm long.   | 4                            |
|    | Awn more than 5.11 mm long. Panicle more than 24.21 mm long. Glumes more than 7.51 mm long. Lateral lobes of the lemma more than 2.43 mm long.   | 6                            |
| 4. | Hairs in upper row on the lemma up to 1.36 mm long.<br>Up to 4 tufts of hairs in upper row of the lemma body.<br>Awn length up to 1.85 times the body of the lemma.<br>Up to 5 spikelets. Up to 3 tufts of hairs in the lower row on the lemma.                                  | 5                            |
|    | Hairs in upper row on the lemma more than 1.36 mm in length. More than 4 tufts of hairs in upper row of the lemma body. Awn length more than 1.85 times the body of the lemma. More than 5 spikelets. More than 3 tufts of hairs in the lower row on the lemma.                  |                              |
| 5. | Palea more than 1.2 times as long as the body of the lemma.<br>Body of the lemma up to 2.6 mm long. Ligule cilia up to 0.32 mm long.   |                              |
|    |  | <i>Danthonia pauciflora</i>  |
|    |  | <i>Danthonia nivicola</i>    |
|    | Palea up to 1.2 times as long as the body of the lemma.<br>Body of the lemma more than 2.6 mm long. Ligule cilia more than 0.32 mm long.   |                              |
|    |  | <i>Danthonia nitens</i>      |
| 6. | Up to 2 tufts of hair in the lower row on the lemma.<br>Hairs in upper row up to 1.07 mm long.<br>Hairs in lower row up to 0.76 mm long. Awn up to 7.66 mm long. Callus hairs up to 0.75 mm long. Palea more than 1.34 times the length of the body of the lemma                 |                              |
|    |  | <i>Danthonia nudiflora</i>   |
|    | More than 2 tufts of hair in the lower row on the lemma.<br>Hairs in upper row more than 1.07 mm long.<br>Hairs in lower row more than 0.76 mm long. Awn more than 7.66 mm long. Callus hairs more than 0.75 mm long. Palea up to 1.34 times the length of the body of the lemma | 7                            |
| 7. | Awn up to 2.06 times the length of the lateral lobe of the lemma. Upper row of hairs placed in the upper third of the body of the lemma. Lateral lobe of the lemma more than 4.2 mm long.  | 8                            |
|    | Awn more than 2.06 times the length of the lateral lobe of the lemma. Upper row of hairs placed in the lower two thirds of the body of the lemma. Lateral lobe of the lemma up to 4.2 mm long.   |                              |
|    |  | <i>Danthonia dimidiata</i>   |
| 8. | Culm scabrous or pilose below the panicle. Up to 5 florets.<br>Ratio of the awn exerted to the total length of the awn up to 0.37. Ratio of the length of the lateral lobe to the body of the lemma more than 1.58.  |                              |
|    |  | <i>Danthonia penicillata</i> |

- Culm glabrous below the panicle. 6 or more florets.  
Ratio of the awn exerted to the total length of the awn  
more than 0.37. Ratio of the length of the lateral lobe  
to the body of the lemma up to 1.58. *Danthonia racemosa*
9. Awn up to 1.59 times the length of the body of the lemma.  
More than 5 nerves extending more than half way  
up the glume. *Danthonia carphoides* var. *angustior*
- Awn more than 1.59 times the length of the body of the  
lemma. Up to 5 nerves extending more than half  
way up the glume. 10
10. Awn up to 5.65 mm long. Panicle up to 23.14 mm long.  
Glumes up to 6.85 mm long. Lateral lobe of the lemma  
up to 2.86 mm long. Body of the lemma up to 2.22 mm  
long. *Danthonia pauciflora*
- Awn more than 5.65 mm long. Panicle more than 23.14  
mm long. Glumes more than 6.85 mm long.  
Lateral lobe of the lemma more than 2.86 mm long.  
Body of the lemma more than 2.22 mm long. 11
11. Callus more than 0.55 mm long. Up to 7 tufts of hairs  
in the upper row on the lemma. Ratio of the awn  
exserted to the total length of the awn up to 0.29.  
Hairs in the upper row on the lemma up to 2.67 mm  
long. Callus hairs up to 0.98 mm long. Up to 9  
spikelets. *Danthonia racemosa*
- Callus up to 0.55 mm long. More than 7 tufts of hairs  
in the upper row on the lemma. Ratio of the awn  
exserted to the total length of the awn more than 0.29.  
Hairs in the upper row on the lemma more than  
2.67 mm long. Callus hairs more than 0.98 mm long.  
More than 9 spikelets. 12
12. Panicle up to 40.65 mm long. Upper row of hairs placed  
in the lower three quarters of the body of the lemma.  
*Danthonia fortuneae-hibernae*
- Panicle more than 40.65 mm long. Upper row of hairs  
placed in the upper quarter of the body of the lemma. 13
13. More than 4 florets per spikelet. Marginal tuft of the  
ligule consisting of less than 10 hairs, the longest up  
to 1.47 mm. Up to 11 tufts of hairs in the upper row  
on the lemma. Hairs in the lower row on the lemma  
up to 0.8 mm long. Ratio of column length to bristle  
length of the awn more than 0.28. *Danthonia semiannularis*
- Up to 4 florets per spikelet. Marginal tuft of the  
ligule consisting of 10 or more hairs, the longest more  
than 1.47 mm. More than 11 tufts of hairs in the upper  
row on the lemma, or a continuous row of hairs  
present. Hairs in the lower row on the lemma more  
than 0.8 mm long. Ratio of column length to bristle  
length of the awn up to 0.28. *Danthonia gracilis*
14. Up to 6 tufts in the upper row of hairs on the lemma. 15
- More than 6 tufts of hairs in the upper row on the  
lemma, or a continuous row of hairs present. 17
15. Culm glabrous below the panicle. Ratio of the length of

- the lateral lobe to the body of the lemma up to 1.58.  
Callus more than 0.92 mm long. More than 6 florets. *Danthonia racemosa*
- Culm scabrous or pilose below the panicle. Ratio of the length of the lateral lobe to the body of the lemma more than 1.58. Callus up to 0.92 mm long. Up to 6 florets. 16
16. Ligule cilia up to 0.39 mm long. Awn length up to 3.09 times the body of the lemma. Ratio of the awn exerted to the total length of the awn up to 0.35. Awn up to 12.53 mm long. Ligule hairs up to 2.03 mm long. Up to 5 florets. Ratio of the length of the lateral lobe to the body of the lemma up to 1.97. *Danthonia penicillata*
- Ligule cilia more than 0.39 mm long. Awn length more than 3.09 times the body of the lemma. Ratio of the awn exerted to the total length of the awn more than 0.35. Awn more than 12.53 mm long. Ligule hairs more than 2.03 mm long. More than 5 florets. Ratio of the length of the lateral lobe to the body of the lemma more than 1.97. *Danthonia pilosa*
17. Upper row of hairs on the lemma body in distinct tufts. Awn length more than 1.28 times the body of the lemma. Awn up to 8.61 mm long. 18
- Upper row of hairs on the lemma body not in distinct tufts. Awn length up to 1.28 times the body of the lemma. Awn more than 8.61 mm long. *Danthonia geniculata*
18. Glume up to 9.25 mm long. Awn up to 8.85 mm long. Up to 3 nerves extending more than half way up the glume. *Danthonia fortuneae-hibernae*
- Glume more than 9.25 mm long. Awn more than 8.85 mm long. More than 3 nerves extending more than half way up the glume. 19
19. Ratio of flat length to total length of lateral lobe of lemma up to 0.37. Hairs in the upper row on the lemma body up to 3.55 mm long. Glumes up to 2.79 mm wide. *Danthonia setacea*
- Ratio of flat length to total length of lateral lobe of lemma more than 0.37. Hairs in the upper row on the lemma body more than 3.55 mm long. Glumes more than 2.79 mm wide. 20
20. Up to 2 tufts of hairs in the lower row on the lemma. *Danthonia laevis*
- More than 2 tufts of hairs in the lower row on the lemma. 21
21. Body of the lemma more than 5.19 mm long. Ratio of floret to glume size more than 0.5. Callus hair up to 1.84 mm long. *Danthonia procera*
- Body of the lemma up to 5.19 mm long. Ratio of floret to glume size up to 0.5. Callus hair more than 1.84 mm long. 22
22. Top row of hairs on the lemma in the lower three quarters of the lemma body. Callus up to 0.5 mm long. Lateral lobe up to 7 mm long. Panicle up to 45 mm long. *Danthonia diemenica*

- Top row of hairs on the lemma in the upper  
quarter of the lemma body. Callus more than  
0.5 mm long. Lateral lobe more than 7 mm long.  
Panicle more than 45 mm long. 23
23. Awn up to 13.17 mm long. Up to 14 spikelets.  
Panicle up to 58.24 mm long. Awn up to 1.58 times  
the length of the lateral lobe of the lemma. Awn  
length up to 3.55 times the body of the lemma. *Danthonia tenuior*
- Awn more than 13.17 mm long. More than 14  
spikelets. Panicle more than 58.24 mm long.  
Awn more than 1.58 times the length of the lateral  
lobe of the lemma. Awn length more than 3.55 times  
the body of the lemma. *Danthonia caespitosa*

Figure 29 — Selecta-key *Danthonia* key.

The results obtained when test data was submitted to this key  
are summarised in Table 18.

Correctly Classified	Incorrectly Classified	Unable to Classify
82%	17%	1%

Table 18 — Classification rate, Selecta-key *Danthonia* key.

It may be noted that the rate of correct identification is the  
best so far.

6.3 Computer time used.

It was expected that the randomisation tests would take  
substantially longer than the Selecta-key tests. This proved to be  
the case in practice.<sup>1</sup> It was expected that representing the data  
in the summarised form used by Selecta-key would lead to a  
reduction in computer time spent in calculation, compared with  
TL-based methods. Table 19 summarises the results of

<sup>1</sup>No exact times are given for either these (or the neural net methods), as timings  
for both were effected by other users running on the Sun at the time, however run  
times were very approximately two orders of magnitude longer than the Selecta-  
key run times, (roughly in the same order of magnitude as the neural net times).  
The times also depend on the level of confidence required that the sample taken  
is representative of the whole sample space, and whether a check is required that  
the exact requirements of the binomial distribution are met, (see discussion  
section 3.1.3.2.3.1 of this thesis). Use of the transportable Pascal system also  
adds a significant time penalty, (compared with running in the installation's  
native version of Pascal). Worst-case running times with the *Acaena* and  
*Danthonia* data, assuming (invalidly) that all characteristics were not normal  
distributions, typically extended beyond minutes into hours for the complete  
run.

measurements of comparative running times of the two algorithms on the *Acaena*, *Danthonia* and storm data (to three significant figures).<sup>1</sup> If the Select-key implementation was coded in Sun Pascal, (and not the transportable system), it is likely that Selecta-key run times would be reduced.<sup>2</sup>

Data	Selecta-key	ID3
<i>Acaena</i>	1.00	10.5
<i>Danthonia</i>	8.01	54.5
storm	0.0237	1510.0

Table 19

Relative running times of the Selecta-key and ID3 algorithms.

## 6.4 Statistical-only versus statistical plus randomisation runs

Theoretically, there is no justification in using Selecta-key's fast data-summarising parametric assumption that the data is normally distributed if the null hypothesis (that there is no difference between the data distribution and a sample drawn from a normal population) can be rejected.

However in practice, the central limit theorem proved very powerful, and in case of both the *Acaena* and *Danthonia* data the keys produced by the (in these cases invalid) assumption that all the characteristics are normally distributed is virtually the same as that produced using the slower randomisation tests where the data characteristic was shown not to be normally distributed.

---

<sup>1</sup>Timings were of a single decision level, repeated a varying number of times in loops, and averaged. Timings were measured in user milliseconds on a Sun 4 which had no other active users at the time the measurements were taken. The storm data is a collection of meteorological data observed before the start of a storm. The conclusion are only two (storm/no storm); the data is unbalanced in the sense that most of the data leads to a "no storm" conclusion. The data contains over three and a half thousand observations, forty-four characteristics.

<sup>2</sup>All real number handling in the ID3 implementation is performed using type Real numbers. As a matter of principle, the transportable system attempts to use real numbers as near to IEEE standard 754 as is reasonable in the particular implementation of Pascal being used. Hence in Selecta-key all real numbers were of type LongReal. Numbers of type LongReal (used in Selecta-key) could be expected to take longer to be processed in an otherwise equivalent calculation than the numbers of type Real (used in the implementation of ID3). Hence if there is any bias in these results, it is likely to be against the Selecta-key method.

The differences in practice were confined to a few minor additional characteristics, and did not alter any of the main decisions on the characteristics to be used for splitting points.

In the case of the complete *Danthonia* data, the null hypothesis that the data distributions of the characteristics are the same is almost always strongly rejected in the cases chosen as splitting points, and when the null hypothesis regarding the splitting points is rejected at the 0%<sup>1</sup> level, the randomisation and statistical tests are the same to within 1 percentage point 92% of the time.<sup>2</sup> This is probably because there is a larger number of examples per characteristic in the case of this data.

The result is less good when there is a smaller number of data per characteristic, as in the *Acaena* data, and even though the key produced by the statistical-only runs of Selecta-key were virtually the same as those produced when the randomisation tests were used with this data, this may not always be the case, and caution is advised in using the results of statistical-only runs when the number of data per characteristic is small. In this case a run using the randomisation tests is advised.

However, even with this caution in mind, the results obtained on this and other data suggest that the decision to do a preliminary run on new data using the statistical assumptions only would not be an unwise one, considering the saving in run time. If the botanical data used so far is to be any guide, the result is likely to be substantially the same as the result of a complete, slower run. However a full check run is advised before the key is accepted.

## 6.5 Alternative Methodologies; Implementation And Test Runs

This section discusses alternative methodologies which may be used in the identification of botanic species and taxa. Most do not allow the production of a portable paper-based identification

---

<sup>1</sup> i.e. <0.5% in this case.

<sup>2</sup> It will be noted that both the statistical and randomisation tests indicate if the 'best' splitting point available is a reasonable one to adopt, given the data being used. This is not the case with some other methods, where a minimisation of some function is used to choose the splitting point, and little regard is given to the reasonableness of the splitting point chosen.

aid suitable for use in the field, but they do provide a comparative baseline which a paper-based methodology (such as a key produced by the Selecta-key process) should at least approach in accuracy.

Four broad groups of alternative methodologies were examined. Section 6.5.1 examines a dozen cluster methodologies. Section 6.5.2 notes the results obtained from two neural net architectures. Section 6.5.3 introduces a simplified derivative of the Selecta-key process, the voting methodology. Section 6.5.4 applies the statistical technique of discriminant analysis to the data sets. Section 6.5.5 summarises the results of these methodologies.

### 6.5.1 Alternative Methodologies — Cluster Analysis

A methodology which could be used for classifying botanic species or taxa is cluster analysis. This methodology can be used when the identification of the specimens is not known.<sup>1</sup> It attempts to cluster specimens into similar groups.

Section 6.5.1.1 outlines the purpose of cluster analysis. Section 6.5.1.2 reports results of tests designed to see if there is a 'natural' number of clusters in the *Acaena* and *Danthonia* data. Section 6.5.1.3 presents summarised rates of identification obtained by applying cluster analysis to these sets of data. Section 6.5.1.4 summarises the results of this investigation of the use of cluster analysis. These sections summarise the investigations of this methodology. A more detailed discussion may be found in Appendix A.

#### 6.5.1.1 Purpose of Cluster Analysis

The purpose of cluster analysis is to place objects into groups or clusters suggested by the data, not defined *a priori*, such that the objects in a given cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar.<sup>2</sup>

---

<sup>1</sup>In this case data translated from the *Acaena* and *Danthonia* data sets did not include species information in a form that was useable to the clustering methodologies.

<sup>2</sup>SAS Institute Inc, *SAS/STAT User's Guide, Release 6.03*, SAS Institute Inc., Cary, NC, 1988, p. 47. The various clustering methodologies employed were



Ideally for the purposes of the identification of botanical species or taxa, each cluster would be composed of only one species or taxa, and the number of clusters would equal the observed number of species or taxa.

#### 6.5.1.2 *Natural Number of Clusters*

The methodology of cluster analysis allows the identification of clusters within the data without reference to the original classification of the data. As mentioned above, this methodology may produce a number of clusters equal to the number of species or taxa, each cluster consisting of one species or taxa. Another possibility is that the number of clusters that were obtained would consistently indicate that a different grouping of species or taxa is appropriate, i.e. a 'natural' number of clusters in the data would be found which was not the same as the species or taxa classification.<sup>1</sup>

Over 30 runs were made with the *Acaena* and *Danthonia* data, using different clustering methodologies. No consistent 'natural' number of clusters was found. As an example, in the case of the *Danthonia* data, results were obtained suggesting the existence of 1, 4, 7, 17, 32 or 100 'natural' clusters, the number depending on the method of analysis employed; (the *Danthonia* data contained 19 species).<sup>2</sup>

#### 6.5.1.3 *Rate of identification using Cluster Analysis*

Since there did not seem to be any 'natural' number of clusters in the data, it was decided to attempt to 'allocate' the clusters to a particular taxa or species.<sup>3</sup> The summarised results of this exercise are given in Tables 20 and 21.

---

implemented in the SAS statistical package, run on a Sun 4 computer. For further details of the SAS package clustering procedures, see Chapter 4 of SAS Institute Inc, *SAS/STAT User's Guide, Release 6.03*, SAS Institute Inc., Cary, NC, 1988.

<sup>1</sup>However, given that the species were chosen by employing cladistic methodologies, this could reasonably be considered unlikely.

<sup>2</sup>This result, together with issues relating to the 'natural' number of clusters in the data, is discussed in greater detail in section A.2.2 of Appendix A of this thesis.

<sup>3</sup>The method used to allocate the clusters to a species or taxa is detailed in the last paragraph of section A.2.2 of Appendix A.

Table 20 summarises the discussion in section A.2.3 and the results presented in Tables 35 to 40 of Appendix A.

Table 21 summarises the discussion in section A.2.4 and the results presented in Tables 41 to 45 of Appendix A.

#### 6.5.1.4 Summary of results using Cluster Analysis

The clustering methodologies employed produced rates of identification superior to that achievable on average by chance for both the *Acaena* and *Danthonia* data. In most cases the rates of identification were also superior to the rate of identification which could be obtained if one had knowledge of the frequency distribution of specimens per species in the data.<sup>1</sup>

The rate of identification achieved in the case of the (complete) *Danthonia* data, although mostly lower in numerical terms than the (incomplete) *Acaena* data results,<sup>2</sup> is proportionally better than the rate noted for the *Acaena* data if one takes note of the expected chance identification.<sup>3</sup> This would appear to confirm the difficulty incomplete data caused to the clustering methodology used.

---

<sup>1</sup>  $5\frac{1}{4}\%$  in the case of the *Danthonia* data; 9% in the case of the *Acaena* data, if the data contained the same number of specimens per species in each data set. However this was not the case in either of these sets of data. If the user had had a knowledge of the number of specimens identified as belonging to each species, the user could have 'guessed' the percentage of specimens belonging to the largest group of species. If this had been the case, the user could have guessed 9.6% for the *Danthonia* data, 23% for the *Acaena* data.

<sup>2</sup> Although in one case the *Danthonia* "identification" rate actually exceeds the corresponding *Acaena* rate, (two-stage density clustering, K=4, see Tables 37 and 44), and is equal in another case, (McQuitty clustering, see Tables 40 and 43).

<sup>3</sup> See the discussion in section A.2.3 and section A.2.4 of this thesis.

Method of Clustering	Correct Identification Rate
Density Analysis, K=2	53% <sup>‡</sup>
Density Analysis, K=3	30%
Density Analysis, K=4	30%
Two-stage Density Linkage, K=2	53% <sup>‡</sup>
Two-stage Density Linkage, K=3	30%
Two-stage Density Linkage, K=4	30%
Single Linkage	32%
Wards	43%
Wards - 10% outliers	45%
Average	38%
Complete	43%
Centroid	30%
EML	43%
Flexible	40%
McQuitty	30%
Median	36%

Table 20 — Rate of identification of *Acaena* data using Cluster Analysis.

<sup>‡</sup> 13 clusters used, not 10 as was the case of all except one of the rest of the results in this table. The higher number of clusters would be expected to result in a higher identification rate, (see the discussion in section A.2.3, especially the paragraph immediately preceding Table 36 in Appendix A).

<b>Method of Clustering</b>	<b>Correct Identification Rate</b>
Density Analysis, K=4	20%
Two-stage Density Linkage, K=3	25%
Two-stage Density Linkage, K=4	45%
Two-stage Density Linkage, K=5	35%
Single Linkage	21%
Wards	36%
Wards - 10% outliers	38%
Average	28%
Complete	29%
Centroid	25%
EML	35%
Flexible	37%
McQuitty	30%
Median	33%

Table 21 — Rate of identification of *Danthonia* data using Cluster Analysis.

Between methodologies, the rate of identification varies widely, from 20% to 53%, and the clustering methodology which gave the equal highest rate with the *Acaena* data (density, 53%) gave the lowest rate on the *Danthonia* data (density, 20%; although different K values, 2 & 4 respectively, were used). In both cases, the most appropriate clustering methodology would appear to be dependent on the multi-dimensional "shape" of the clusters which naturally occur in the data.

In summary, clustering methodology would appear to be a useful, albeit limited methodology in the classification of botanical species and taxa.

### 6.5.2 Alternative Methodologies — Neural Nets.

Neural net methodology originated from attempts to emulate the functioning of the human brain. This methodology has been used for classification tasks, and was investigated as an alternative to the Selecta-key methodology.<sup>1</sup>

After investigating several types of neural net, the multi-layer perceptron net was chosen.<sup>2</sup> Two types of multi-layer nets were used. Section 6.5.2.1 discusses the results obtained with a neural net employing the Aristotelian assumption of complete enumeration. Section 6.5.2.2 discusses the results obtained with a more orthodox neural net architecture which dismisses the Aristotelian assumption and also allows real-valued input variables to be used. Section 6.5.2.3 summarises the experiences attained using these neural net methodologies for the species identification task.

#### 6.5.2.1 Species Identification, Aristotelian neural net

The first neural net written used the Aristotelian assumption of complete enumeration.<sup>3</sup> A typical average result produced by this type of net are presented in Table 22.<sup>4</sup>

Correctly Classified	Incorrectly Classified	Unable to Classify
47.9%	44.6%	7.5%

Table 22 — Classification Rate, Aristotelian Neural Net method using *Danthonia* Data,

Although these results are comparable with the best of the cluster analysis results, the Aristotelian requirement of complete enumeration produced problems in practice, namely:-<sup>5</sup>

<sup>1</sup>This methodology, and some of the theory behind it, is discussed in much greater detail in Appendix B of this thesis.

<sup>2</sup>See Figure 38 and the discussion in section B.3.7 of Appendix B of this thesis.

<sup>3</sup>For further details see section 1.1.1.2 in the main body of this thesis, and section B.6.2.2 and Table 46 in Appendix B of this thesis.

<sup>4</sup>Table 22 is obtained by averaging the results presented in Table 46 of Appendix B of this thesis.

<sup>5</sup>For a fuller discussion, see section B.6.2.2 of Appendix B of this thesis.

- a) The Aristotelian requirement of complete enumeration meant that the input to the neural net had to be categoric. This meant that the real-valued measurements of the specimens had to be categorised.
- b) The accuracy of identification of the botanical species was heavily dependent on the choice of the categorisation points, and an indication of the desirable location of categorisation points was not inherent in the methodology.
- c) Categorisation could cause duplicate patterns to occur between species in either or both the test and learning data.
- d) The 80% training/20% test data regime could cause a violation of the Aristotelian assumption that all of the 20% test patterns were previously observed in the 80% learning data.<sup>1</sup>

These problems were sufficiently discouraging to lead to a decision to halt work on this prototype software. Whilst producing good results with tasks such as character recognition, it proved to have too many undesirable features when it was applied to the identification of botanic species & taxa. It was therefore decided to commence work on a prototype neural net which could handle real-valued input, generalise, and hopefully produce better results when applied to the species and taxa identification task.<sup>2</sup>

#### *6.5.2.2 Species Identification, non-Aristotelian neural net*

Software to implement a multi-level perceptron net had been written and was in it's initial testing stages when the versatile MITRE neural net simulator with it's excellent graphical user interface became available.<sup>3</sup> This simulator was adopted for use in

---

<sup>1</sup>This problem is similar to the problem noted in section 5.4 of this thesis, but is made worse by the categorisation necessary to meet the Aristotelian assumption.

<sup>2</sup>Zeidenberg comments 'Without the ability to generalise, neural network models would be like look-up tables, which are not very interesting', see Zeidenberg, Matthew, *Neural Networks in Artificial Intelligence*, Ellis Horwood, New York, 1990, p. 17.

<sup>3</sup>See Leighton, R., and Wieland, A., *The Aspirin/MIGRAINES Software Tool User's Manual, Release 4.0*, The MITRE Corporation, Washington, 1991.

these trials, and further development of the author's locally written neural net software ceased, as there seemed to be no point in re-inventing the wheel.

The MITRE simulator was used to set up a three-level perceptron net, using sigmoid functions in the hidden layer.<sup>1</sup> A program was written to translate the data from the Selecta-key format into the format required by the MITRE neural net simulator.<sup>2</sup> The missing values in the *Acaena* data presented problems, which were overcome by having the data conversion program produce synthetic data which was incorporated into multiple copies of the *Acaena* data.<sup>3</sup> The *Danthonia* data was complete, and so options in the data translation program were set to produce just one copy of the *Danthonia* data in the translated data. Multiple runs were obtained with approximate stratified split 80% learning/20% test sets of each data.<sup>4</sup> Average results for the *Acaena* data are given in Table 23.<sup>5</sup>

Network Configuration	Correctly Classified	Incorrectly Classified	Unable to Classify
31 - 63 - 11	60%	12%	28%
31 - 41 - 11	58%	10%	32%
31 - 21 - 11	56%	12%	32%
31 - 11 - 11	55%	10%	35%

Table 23 — Average Classification Rate, non-Aristotelian neural net method using *Acaena* Data.

The first, second and third numbers in the "Network Configuration" column of Tables 23 and 24 refer to the number

<sup>1</sup>For further information about the use of the sigmoid function, see section B.4 of Appendix B of this thesis, particularly the areas around Figures 42 and 45. For further information on the three-layer perceptron net, see section B.3.7 of Appendix B.

<sup>2</sup>For more information about the conversion programs, see section 4.7 b) & f) of this thesis. Although the production of these programs took more time than completing the local neural net simulator, the overall result was a more versatile package.

<sup>3</sup>For further information on the synthetic data, see sections B.6.1.2 and B.6.2.1 of Appendix B of this thesis.

<sup>4</sup>For an explanation of 'approximate stratified split', see section 5.4 of this thesis.

<sup>5</sup>These are summary results. For more detailed results, see Tables 47 to 51 (and the surrounding text) in Appendix B of this thesis.

of network nodes in the input, hidden and output layers respectively.<sup>1</sup>

The average results obtained for the *Danthonia* data are presented in Table 24.<sup>2</sup> Again these results are averages obtained from multiple runs which were obtained with data split by the approximate stratified split method into 80% learning/20% test sets of (in this case) the *Danthonia* data.<sup>3</sup>

Network Configuration	Correctly Classified	Incorrectly Classified	Unable to Classify
41 - 83 - 19	56%	5%	39%
41 - 63 - 19	50%	8%	42%
41 - 43 - 19	47%	8%	45%
41 - 23 - 19	48%	11%	41%
41 - 13 - 19	47%	17%	36%
41 - 8 - 19	42%	16%	42%

Table 24 — Average Classification Rate, Neural Net method using *Danthonia* Data.

It will be noted that the non-Aristotelian net, at best, obtained higher correct recognition rates than the Aristotelian net. However the training times were three to four orders of magnitude greater for the non-Aristotelian net.

### 6.5.2.3 Neural net summary.

It can be seen from Tables 22 to 24 that the classification rates obtained are well above those which would have been obtained by chance.<sup>4</sup> They are also generally superior to the rates

<sup>1</sup>The number of hidden nodes varies because, although Hecht-Nielsen's re-statement of Kolmogorov's proof of the thirteenth theorem of Hilbert indicates an adequate number of middle-level neurons, it does not specify the minimum number of hidden-level nodes necessary for a seamless mapping of input to output level nodes. See the discussion in section B.4 of Appendix B of this thesis.

<sup>2</sup>For more detailed results, see Figures 52 to 58 of Appendix B of this thesis.

<sup>3</sup>For an explanation of 'almost random' see section 5.4 of this thesis.

<sup>4</sup> $5\frac{1}{4}\%$  in the case of the *Danthonia* data; 9% in the case of the *Acaena* data, if the data contained the same number of specimens per species in each data set. However this was not the case in either of these sets of data. If the user had had a knowledge of the number of specimens identified as belonging to each species,



obtained by use of clustering methodologies. The results obtained by the non-Aristotelian neural net simulation were also generally superior to the results obtained by the Aristotelian neural net in this task of species identification, this being particularly so if one notes that the former does not require the estimation of splitting points.

### 6.5.3 Alternative Methodologies — Voting

An alternative methodology which could be used for the task of species or taxa identification was developed as a simplified offshoot of the Selecta-key methodology. It was called the Voting Method. Section 6.5.3.1 briefly discusses the methodology and it's implementation. Section 6.5.3.2 presents the results obtained by use of this methodology. Section 6.5.3.3 discusses the results obtained by the methodology. More detail of the Voting methodology may be found in Appendix C of this thesis, where the methodology is discussed in greater detail than in the following brief sections.

#### 6.5.3.1 Discussion of Voting Methodology

In this methodology, the data is split into training and test data.<sup>1</sup> Measurements observed for each characteristic of each species of the training set of data are grouped, and the groups ranked for each characteristic. Splitting points are established for each species per characteristic.

Identification of specimens in the test data can then be made by comparing the measurement for each characteristic with the "template" established from the training data; each species receiving a "vote" if the specimen's characteristic measurement falls within the species' splitting points for that characteristic. The species with the highest "vote" total is declared to be the likely species to which the specimen belongs.<sup>2</sup>

This methodology was implemented in Pascal 2.0 on a Sun 4 computer, using the portability package developed as a part of

---

the user could have 'guessed' the percentage of specimens belonging to the largest group of species. If this had been the case, the user could have guessed 9.6% for the *Danthonia* data, 23% for the *Acaena* data.

<sup>1</sup>For more detail see section 5.4 of this thesis.

<sup>2</sup>For more detail, see section C.1.1 of Appendix C of this thesis.

this project. As implemented, the programs only handle cases where the data is complete, and hence results are only presented for the *Danthonia* data.<sup>1</sup>

6.5.3.2 Results obtained by use of the Voting Methodology

Eight runs were obtained from training and test sets selected from the *Danthonia* data.<sup>2</sup> The average results obtained are shown in Table 25.<sup>3</sup>

Correctly Classified	Incorrectly Classified
49.3%	50.7%

Table 25 — Average Classification rate — Voting Methodology (First Choice only).

On many occasions it was noted that the first two choices were close. As an indication of this, Table 26 lists the average classification rate if the correct species occurs within the first two voting choices.<sup>4</sup>

Correctly Classified	Incorrectly Classified
63.8%	36.2%

Table 26 — Average Classification rate — Voting Methodology (First Two Choices only).

6.5.3.3 Summary — Voting Methodology

In both Tables 25 and 26, the classification rate is well above that which could, on average, be obtained by chance.<sup>5</sup> The

<sup>1</sup>Extension of this methodology to include the ability to handle data with missing values is possible, subject to the precautions mentioned in section C.1.2 of Appendix C of this thesis.

<sup>2</sup>For more details on the methods of preparing the data, see section C.2.1 of Appendix C of this thesis.

<sup>3</sup>This Table is similar to Table 60 of Appendix C; the full results from which this average is obtained are listed in Table 59 of Appendix C of this thesis.

<sup>4</sup>This Table is similar to Table 62 of Appendix C; the full results from which this average is obtained are listed in Table 61 of Appendix C of this thesis.

<sup>5</sup> $5\frac{1}{4}\%$  in the case of the *Danthonia* data; if the data contained the same number of specimens per species. However this was not the case. If the user had had a knowledge of the number of specimens identified as belonging to each species,

recognition rate was better than that obtained by use of the clustering methodologies, in the same range as those obtained by use of the neural net methodologies. This classification process (after training has taken place) probably makes less computational demands than any other computer-based method examined in this thesis, (this statement assumes Selecta-key identifications take place via a paper key).

#### 6.5.4 Alternative Methodologies — Discriminant Analysis

Several alternative statistical methodologies which could be used for the task of species or taxa identification are grouped under the general heading of discriminant analyses. Two of these methodologies were employed to examine the *Danthonia* and *Acaena* data, one method making parametric assumptions, the other non-parametric. Section 6.5.4.1 discusses the parametric test and the results obtained from it, section 6.5.4.2 the non-parametric test and its results, and section 6.5.4.3 makes some summary comments about these analyses. These sections are brief, and more detail about these methodologies may be found in Appendix D of this thesis.

##### 6.5.4.1 Parametric Discriminant Analysis

This methodology assumes the distributions of the measurements are multivariate normal.<sup>1</sup> It uses a training set of data to obtain a quadratic discriminant function which may then be used to classify specimens in a set of test data. Section 6.5.4.1.1 comments briefly on the methodology as implemented, and section 6.5.4.1.2 lists the results obtained from this methodology.

##### 6.5.4.1.1 Parametric methodology employed.

The discriminant analysis employing the assumption of multivariate normal distributions per species per characteristic

---

the user could have 'guessed' the percentage of specimens belonging to the largest group of species. If this had been the case, the user could have guessed 9.6% for the *Danthonia* data,

<sup>1</sup>This is not necessarily a valid assumption, see Appendix E, section E.4.1, where tests suggest that about two-thirds of the distributions of each data are non-normal. Thus the success of this test will depend to a large extent on the robustness of the test to the presence of a proportion of non-normal distributions amongst the data.

was available in the SAS package running on a Sun 4 computer. Previously written Pascal program converted the *Danthonia* and *Acaena* data into formats suitable for use with the SAS package. They also provided the 80% training/20% test splits on the basis of an approximate stratified split.<sup>1</sup>

6.5.4.1.2 Results obtained from the parametric discriminant analysis

Eight runs were made with the *Danthonia* data. The average rate of classification obtained is shown in Table 27.<sup>2</sup>

Correctly Classified	Incorrectly Classified
56%	44%

Table 27 — Average Classification rate using the *Danthonia* data

Seven runs were made with the *Acaena* data. The SAS restriction requiring full data for every specimen eliminated about three-quarters of the specimens in both the training and data sets, making the results of the individual runs somewhat erratic.

Table 28 presents the average rate of identification for completely described *Acaena* specimens.<sup>3</sup>

Correctly Classified	Incorrectly Classified
43%	57%

Table 28 — Average Classification rate using the *Acaena* data, (excluding the specimens having incomplete data).

If the rate of identification was defined to include the incompletely described specimens in the test data eliminated by

<sup>1</sup>For more detail on the treatment of the data see sections 5.4 and 4.7 b) & g) in the body of this thesis, and section D.2.1 of Appendix D of this thesis.

<sup>2</sup>Table 27 is similar to Table 64 in Appendix D. For more detail see Table 63 of Appendix D which lists the outcomes of the individual *Danthonia* runs, the results of which were averaged to obtain Table 27.

<sup>3</sup>Table 28 is similar to Table 65 of Appendix D of this thesis, which contains a more complete discussion of these results.

the SAS requirement for complete data, the situation would be as shown in Table 29.<sup>1</sup>

Correctly Classified	Incorrectly Classified	Unable to Classify
9%	12%	79%

Table 29 — Average Classification rate using the *Acaena* data, (including the specimens having incomplete data).

6.5.4.2 *Non-Parametric Discriminant Analysis*

Non-parametric discriminant analyses do not need any assumption that the data fits into a normal distribution, and are often useful in the case where data is grouped into irregular distributions, as is likely to be the case with the data sets being employed in these tests.<sup>2</sup>

6.5.4.2.1 *Non-parametric methodology employed.*

Epanechnikov's kernel method was used to generate a non-parametric density analysis estimate, which was then applied to the test data. This methodology was available in the SAS package which was programmed on a Sun 4 computer.

6.5.4.2.2 *Results obtained from the non-parametric discriminant analysis.*

The average results obtained from eight runs of the *Danthonia* data (80% training/20% test approximate stratified splits) are shown in Table 30.<sup>3</sup>

<sup>1</sup>Table 29 is similar to Table 66 of Appendix D of this thesis, which contains a more complete discussion of these results.  
<sup>2</sup>Many distributions in the *Danthonia* and *Acaena* data sets appear not to be mesokurtic, see section E.4.1 of Appendix E of this thesis.  
<sup>3</sup>Table 30 is similar to Table 68 of Appendix D of this thesis, which contains a more complete discussion of these results, including the results of the individual runs which were averaged to obtain this Table, (see Table 67).

Correctly Classified	Incorrectly Classified	Unable to Classify
74%	25%	1%

Table 30 — Average Classification rate using the *Danthonia* data, Epanechnikov's kernel methodology.

Seven runs were undertaken using the *Acaena* data, but the results were worse than those of Table 29, and are thus not presented here because the restriction on incomplete data implemented in SAS may be the cause of the poor result, rather than any shortcomings in Epanechnikov's methodology.

### 6.5.5 Summary — Alternative Methods

The range of results obtained from the alternative methodologies are summarised in Table 31 for the *Acaena* data Table 32 for the *Danthonia* data.

Method	Correctly Classified	Incorrectly Classified	Unable to Classify
Clustering - completely described specimens only	30 - 53%	47 - 70%	—
Neural Net - Non-Aristotelian	55 - 60%	10 - 12%	28 - 35%
Discriminant Analysis - Parametric assumption - complete specimens only	43%	57%	—
Discriminant Analysis - Parametric assumption - all specimens	9%	12%	79%

Table 31 — Ranges of Classification Rates — *Acaena* data.

Method	Correctly Classified	Incorrectly Classified	Unable to Classify
Clustering	20 - 45%	55 - 80%	—
Neural Net - Aristotelian	48%	45%	7%
Neural Net - Non-Aristotelian	42 - 56%	5 - 17%	36 - 45%
Voting (first choice)	49%	51%	—
Voting (first two choices)‡	64%	36%	—
Discriminant Analysis - Parametric assumption	56%	44%	—
Discriminant Analysis - Non-parametric assumption	74%	25%	1%

Table 32 — Ranges of Classification Rates — *Danthonia* data.

These results will be discussed further in section 6.6.

---

‡ Note that this is not strictly comparable with the rest of the results in this table, as it includes the first two choices, only the first choices being presented in the rest of the methodologies. It was decided to include this when it was noticed that many of the choices were very close, and it was an interesting property of this methodology that ranked choices were available.

## 6.6 Discussion of Results

Tables 33 and 34 presents summaries of all the results obtained for the *Danthonia* and *Acaena* data, respectively.

Method	Correctly Classified	Incorrectly Classified	Unable to Classify
Chance	5%	95%	—
Pre-knowledge of Specimen numbers per Species	10%	90%	—
Clustering	20 - 45%	55 - 80%	—
Neural Net:- Aristotelian	48%	45%	7%
Neural Net:- Non- Aristotelian	42 - 56%	5 - 17%	36 - 45%
Voting (first choice)	49%	51%	—
Voting (first two choices)	64%	36%	—
Discriminant Analysis - Parametric assumption	56%	44%	—
Discriminant Analysis - Non-parametric assumption	74%	25%	1%
Collier's <i>Danthonia</i> Key	70%	12%	18%
Selecta-key	82%	17%	1%

Table 33 — Classification of *Danthonia* data.

Examination of Tables 33 and 34 will show that the clustering methodologies generally gave the worse results. This is to be expected as clustering, by it's very nature, works best with well separated data sets. Botanic data sets, of which the *Acaena* and *Danthonia* data sets are difficult examples, typically contain much data which is poorly separated. For this reason, clustering would generally not be the methodology of choice for identification of botanic specimens.



Method	Correctly Classified	Incorrectly Classified	Unable to Classify
Chance	9%	91%	—
Pre-knowledge of Specimen numbers per Taxa	23%	77%	—
Clustering - complete specimens only	30 - 53%	47 - 70%	—
Neural Net - Non-Aristotelian	55 - 60%	10 - 12%	28 - 35%
Discriminant Analysis - Parametric assumption - complete specimens only	43%	57%	—
Discriminant Analysis - Parametric assumption - all specimens	9%	12%	79%
Collier's Summary Key	63%	15%	22%
Quinlan's C4.5 Algorithm	69%	5%	26%
Orchard's Key	70%	28%	2%
Selecta-key imitation of Orchard's Key	68%	24%	8%
Selecta-key	75%	13%	12%

Table 34 — Classification of *Acaena* data.

The multivariate normal, neural net and voting methodologies all produced reasonable results, in each case well above chance.

The multivariate normal result was somewhat surprising, as the data was generally unfavourable to the assumptions of normality made in these tests. The multivariate normal results with the *Danthonia* data were better than the majority of the clustering and voting (first choice) methodologies, and in the same range as the neural net methodologies. It even obtained a reasonable to good result with the difficult *Acaena* data if one ignored the missing test data, (but an unsatisfactory result if these were included).

The non-parametric tests provided an excellent rate of recognition for the *Danthonia* data, the best of the alternative methodologies. However its rate of identification of the *Acaena* data was unsatisfactory, although it was uncertain if this would also have been the case if the SAS program had not rejected all specimens with incomplete data.

The neural net identification rates are seen to be in a similar range to the multivariate normal results. The non-Aristotelian net, at best, performed better than the Aristotelian net, but used training times which were three to four orders of magnitude greater. The Aristotelian neural net's training times were much greater than the Selecta-key parametric methodology times. This suggests the neural net methods are robust, produce results well above chance, but are very compute intensive.

The voting methodology identification rate will be seen to be below the rate obtained by use of the entropy and Selecta-key methodologies. This methodology, like clustering, is likely to be badly affected by poorly separated data. Its classification process (after training has taken place) probably makes less computational demands than any other computer-based method examined in this thesis, (this statement assumes Selecta-key identifications take place via a paper key). The Voting methodology also has the advantage that, like Selecta-key, it may be delivered in a paper format and does not require the presence of a computer to allow identification to take place. Unlike Selecta-key, it has the disadvantage of being an automatic process, and hence does not allow the Expert to participate in the formation of the representation of the learnt knowledge.<sup>1</sup>

The best *Danthonia* identification rate for the non-key algorithms was obtained using non-parametric discriminant analysis. In the case of the *Acaena* data, this methodology was handicapped by the SAS implementation demanding complete data on all specimens, and this SAS limitation made any SAS-based methodology not a method of choice if the data set contained a significant number of specimens which do not have complete data specified for each characteristic. This SAS

---

<sup>1</sup>These matters are discussed in more detail in sections C.3 and C.4 of Appendix C of this thesis.

limitation, while statistically desirable, is a major drawback in the case of many botanic data sets.

With the notable exception of Quinlan's C4.5 algorithm (see Table 34), the key-construction algorithms generally produced better identification rates for the botanic data sets examined than the competing methodologies. It will be noted that the rate of identification of the *Acaena* data is comparable with that produced by the application of a published key produced by Orchard, who is recognised as a distinguished expert in this field.

A significant factor in the high rate of recognition of the botanic specimens is the richness of the characteristics at each branch of the decision key. Selecta-key makes the choice of additional characteristics easy. This richness is particularly important in the case of botanic data sets, as it makes handling missing and highly variable data easier. Even if some of the data needed for a decision at a decision key splitting point is missing, a rich description allowing a choice on the basis of the minimum Hamming distance will often make a meaningful decision possible.

The effect of this is best illustrated in the case of the *Acaena* data. The SAS elimination of specimens which had missing data made overall identification rates very low. By contrast, methodologies which result in rich keys such as Collier's Danthonia key and the Selecta-key Danthonia key, handled this type of data in a more satisfactory manner. The higher rate of identification achieved by the Selecta-key derived key<sup>1</sup>, compared with Collier's key, is mainly due to an increased richness of this key. Selecta-key's ability to indicate additional favourable characteristics meant that the Selecta-key derived key had approximately 40% more characteristics included than had Collier's key. This reduced the "unable to classify" group by 17%, (of which 12% were correctly and 5% incorrectly identified, a result still above chance).

---

<sup>1</sup>The author wishes to acknowledge that this key was prepared with the help of Collier, who very kindly lent his considerable botanic knowledge to this project.

## 6.7 Summary of Results

The Selecta-key methodology, as implemented, was compared with Collier's implementation of Quinlan's entropy-based methodology ID3 and Quinlan's C4.5, discriminant analysis,<sup>1</sup> various clustering procedures (which in some cases were followed by a canonical discriminant analysis), two implementations of neural net methodologies, and a simpler variation of the Selecta-key methodology referred to as the Voting methodology.

The results obtained indicate that keys produced by use of the Selecta-key interactive inductive inference technique allow the production of an easily-transportable paper key which is likely to allow a rate of identification of botanic species comparable with existing computer-based methodologies. This methodology is considered to be likely to be a particular use in botanic data sets which often contain a significant proportion of specimens for which complete data is not available.

In summary, it is considered that the method of interactive statistical inference can be a useful tool in botanical key construction. In the cases considered, it saves time compared with previous methodologies, and assisted the production of a rich key which can exhibit comparable classificatory power to existing methods in the case of complete data, and a more accurate classificatory power in the case of data sets which contain a significant proportion of incompletely-specified specimens.

---

<sup>1</sup>These methodologies were not compared with Bayesian methods, but the author notes Payne & Preece's comment that, in the only direct comparison between these methodologies known to them, Bayesian and discriminant analysis gave equal accuracy; see: Payne, R. W. and Preece, D. A., 'Identification Keys and Diagnostic Tables: a Review', *Journal of the Royal Statistical Society, Series A*, The Royal Statistical Society, Volume 143, 1980, p. 290.

# Future Work

There are several aspects of the Selecta-key approach which could benefit from more work.

At present the method of selecting alternative characteristics is not ideal. The taxa are listed on the screen in order of increasing numerical size, and since this order is, in general, different for each characteristic, it makes finding the similar splits needed for alternative characteristics tedious. It would be better if the alternative splits were presented in a manner which made them easier for the user to recognise, perhaps grouped together.

The user interface is at present text-based, and treats the screen as a glass teletype. Improvements to the transportable Pascal system would allow screen addressing and screen handling in general to be included in Selecta-key. These improvements have been started, but there is still a fair amount of work yet to be done before this becomes a possibility in the transportable version of Selecta-key.

The randomisation tests are, at present, slow. While they will always be compute intensive, some code optimisation and an investigation of the size of window needed for the approximate randomisation tests could prove useful in reducing run times.

Since it was found that, in practice, there was little difference between the keys produced using the parametric and non-parametric assumptions, it would be useful to have an additional program in the series which took as input a key produced by Selecta-key using all parametric assumptions, and checked the key (using a mix of parametric and non-parametric assumptions as appropriate) in batch mode.

It would be useful to validate the results on other large botanic data sets.

At present there is no manual available for the Selecta-key system.

If these changes were made, the Selecta-key system would be more user-friendly.

# Conclusions

The subject of this thesis is induction, specifically the application of induction and the acceptability of inductively based computer assisted key generation methodologies applied to the botanic area.

For a methodology to be acceptable, the user must be able to accept that the basis of its operations are believable. It has been argued that basic axioms of belief would make the concept of Artificial Intelligence unacceptable to many users. It is therefore considered advisable that the any system employing "Artificial Intelligence" techniques be promoted for reasons other their inclusion in the methodology; in this case the key generation system would be promoted as an aid to the researcher, based on the long history of computing and key generation, which pre-dates the involvement of artificial intelligence in this field.

A second condition for the methodology to be acceptable is that the researcher understand the type of reasoning used in the system. It has been argued that inductive reasoning is understood by a much wider proportion of the human population than deductive reasoning, and thus its use in the methodology is preferable because the Selecta-key system is intended to be acceptable to researchers with expertise in a wide variety of biological fields.

A third condition for acceptability of a methodology is that the user feel comfortable with the result obtained from the methodology. It has been argued that over two decades use of automatic key-generation algorithms have not resulted in some of the researchers feeling comfortable with the results. It is thus argued that an interactive key-generation methodology which combines the best aspects of both elements in the key-generation process (the tireless calculation ability of the computer plus the background knowledge and common sense of the researcher) will result in a key with which both the researchers and users will feel comfortable.

Another aspect of the methodology developed in this thesis is that the key generated by the co-operative effort of the computer

and researcher will only have splits which are statistically acceptable to a level specified by the researcher. If the data is inadequate to support the generation of a complete key, only a partial key will be generated. In many practical biological applications this is regarded as being preferable to automatic key generation methodologies which generate a key for all the data, regardless of whether the resultant key can be statistically supported by the data or not. Again it is argued that the researcher will be more comfortable with a key that can be supported by the data to a specified degree of statistical acceptability.

It is further argued that the fact that the methodology can be used to produce polythetic keys, (instead of the monothetic keys to which some methodologies are limited) is another reason for acceptance by the researcher and user, in that less errors of identification typically occur with the use of a polythetic as opposed to a monothetic key.

Evidence is presented that the inclusion of the researcher in the key generation process means that many difficulties in the area of the description of data from living sources is lessened. The researcher can use his or her background knowledge in choosing characteristics for splitting points in cases where the characteristics used are not completely described in the data (e.g. time-varying, inherently qualitative or difficult-to-measure data).

It is argued that using a co-operative methodology will ease the demands on the researcher, in that specific purposes keys may be generated from the data without the researcher being required to edit the data between runs, (a process necessary for many automatic key-generation systems).

For any proposed methodology to be acceptable, it must produce results which are accurate. An extensive series of comparisons were made between the proposed methodology and several other methodologies, including discriminant analysis, two entropy-based methodologies, multiple runs of twelve clustering methodologies (including some runs under several varying parametric conditions), multiple runs (with different numbers of hidden nodes) of two neural net methodologies,

another simplified and very fast methodology developed during the course of this work, voting, and a paper-based key produced manually by a domain expert. In all cases the methodology proposed in this thesis produced results which were similar or better than the results produced by the competing methodologies. When used with botanic data which was markedly incomplete, the Selecta-key methodology produced a higher rate of correct identification than competing methodologies. With parametric data, the methodology also produced results with a far lower computational load than most of the other methodologies (a notable exception being the "voting" methodology).

If the results of the comparisons are to be regarded as being acceptable, the data used must be typical of the data to be used in practice, not artificial data designed to favour the methodology being proposed. It is also argued that the data used in the comparisons is "real" data obtained from botanic sources, and during the course of these investigations the data was examined with the conclusion that it contained many of the problems which can typical occur in collections of data obtained from botanic sources; in that it was poorly separated, contained outliers, many of the specimens were incompletely described, the distributions of some of the subsets of data were parametric and some not, it contained both continuous and interval data, some of the characteristics were from time-varying portions of the species under consideration, some of the characteristics used were traditionally regarded as being qualitative rather than quantitative, and the difficulty in obtaining the measurements varied widely with the characteristic under consideration. It was thus argued that the results obtained in the extensive series of comparisons is valid.

In summary, it is argued that the methodology proposed in this thesis is a practical and useful methodology that has been used to produce understandable keys of excellent quality from real data of botanic origin whilst imposing a reasonable and acceptable load on both the computer and domain expert involved.



# Reference List

Aleksander, Igor, *Designing Intelligent Systems*, Billing & Sons Limited, Worcester, Great Britain, 1984.

Alexander, James, *Intelligence: Natural and Artificial*, Seminar handout, Hobart, Tasmania, 15 October 1992.

Ali, A. M., 'Probability - Uncertainty - Simulation', in Jelen, F. C., (Ed.), *Cost and Optimization Engineering*.

Alkon, Daniel L., *Memory Storage and Neural Systems*, Scientific American, July 1989, pps. 26 - 34.

Alkon, D.L., Blackwell, K.T., Barbour, G.S., Rigler, A.K. and Vogl, T.P., *Pattern-Recognition by an Artificial Network Derived from Biologic Neuronal Systems*, Biological Cybernetics, No. 62, pps. 363-376, 1990.

Almuallim, Hussein and Dietterich, Thomas G., 'On Learning More Concepts', in Sleeman, Derek and Edwards, Peter (Eds.), *Machine Learning: Proceedings of the Ninth International Workshop*, Morgan Kaufmann Publishers, San Mateo, 1992.

Alty, J.L., & Coombs, M.J., *Expert Systems, Concepts and Examples*, NCC Publications, Manchester, 1984.

Amari, Shun-Ichi, *Field Theory of Self-Organising Neural Nets*, IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-13, No. 5, September/October 1983.

Anderson, Ian, "AI is start naked from the ankles up", *New Scientist*, 15 November 1984, IPC Magazines Ltd., England, 1984.

Anderson, James A., *Cognitive and Psychological Computation with Neural Models*, IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-13, No. 5, September/October, 1983

Anderson, Mike, *Intelligence and Development A Cognitive Theory*, Blackwell Publishers, Oxford, 1992.

Andreas, Burton G., *Experimental Psychology*, John Wiley and Sons, Inc., New York, 1960.

Aoki, Chiye and Siekevitz, Phillip, *Plasticity in Brain Development*, Scientific American, December 1988.

Augustine, St., 'The Freedom of the Will', in Berofsky, Bernard (Ed.), *Free Will and Determinism*.

Aune, Bruce, *Knowledge of the External World*, Routledge, London, 1991.

Bach, Ivan N., *Data Complexity*, Neuron-Digest, Vol. 5, No. 51, December 1989.

Bacon, Francis, *First Book of Aphorisms*, quoted by Forsyth, R. S. in 'The Evolution of Intelligence', *The Third International Expert Systems Conference*, Learned Engineering, Oxford, 1987,

Bagnall, Diana, 'New crimes of the times', *The Bulletin with Newsweek*, Vol. 114, No. 5843, ACP Publishing Pty. Ltd., Sydney, 3 November 1992.

Bala, Jerry W., Michalski, Ryszard S. and Wnek, Janusz, 'The Principal Axes Method for Constructive Induction', in Sleeman, Derek and Edwards, Peter, *Machine Learning Proceedings of the Ninth International Workshop*, Morgan Kaufmann Incorporated, 1992.

Balzer, R.M., 'Search for a Solution: A case study', in Walker, Donald E. and Norton, Lewis M. (Eds.), *Proceedings of the International Joint Conference on Artificial Intelligence*, Washington, 1969.

Barlow, H.B., *Single units and sensation, a neuron doctrine for perceptual psychology?*, Perception 1, 1972, pps. 371-394, (not seen, referred to in Rolls, Edmund, "The Representation and Storage of Information in Neuronal Networks in the Primate Cerebral Cortex and Hippocampus" in: Durbin, Richard, Miall, Christopher and Mitchison, Graeme (Eds.), *The Computing Neuron*, Addison-Wesley, England, 1989).

Barr, Murray L. and Kiernan, John A., *The Human Nervous System*, (fourth edition), Harper and Row, Philadelphia, 1983.

Bee, Helen, *Social Issues in Developmental Psychology*, Harper and Row, New York, 1978.

Bee, Helen, *The Developing Child*, Harper & Row, New York, 1978.

- Begley, Sharon & Ramo, Joshua Cooper, "Not just a pretty face", *The Bulletin with Newsweek*, 2 November 1993, ACP Publishing Pty. Ltd., Sydney, 1993.
- Berofsky, Bernard (Ed.), *Free Will and Determinism*, Harper and Row, New York, 1966
- Beth, Evert W. & Piaget, Jean, *Mathematical Epistemology and Psychology*, D. Reidel Publishing Company, Dordrecht, Holland, 1966.
- Beynon, David, *Father of AI blasts the 'philosophers'*, Computerworld, September 6, 1991.
- Bicking, C. A., 'Process Control by Statistical Methods', in Juran, J. M., Gryna, Jr., Dr. Frank M., Bingham, Jr., R. S., (Eds.), *Quality Control Handbook*.
- Biller, Henry and Meredith, Dennis, *Father Power*, David McKay Company. Inc., New York, 1975.
- Blakemore, Colin, 'Computational Principles of the Visual Cortex', *The Psychologist*, Vol. 4, No. 2, February 1991.
- Blakemore, Colin, *Mechanisms of the Mind*, Cambridge University Press, Cambridge, 1977.
- Block, H.D., *The Perceptron, A Model for Brain Functioning*, in Review of Modern Physics, 34(1), January 1962.
- Bloomfield, Brian P., 'Capturing expertise by rule induction,' in *The Knowledge Engineering Review*, Cambridge University Press, Vol.2, No. 1, March 1987.
- Boden, Margaret A., 'Real World Reasoning', in Cohen, L. Jonathon, & Hesse, Mary, (Ed.), *Applications of Inductive Logic*.
- Boden, Margaret A., *Artificial Intelligence and Natural Man*, Basic Books, Inc., New York, 1977.
- Boden, Margaret A., *Artificial Intelligence in Psychology: Interdisciplinary Essays*, Bradford Books, MIT Press, Cambridge, U.S.A., 1989, (not seen), quoted in Massaro, Dominic W., Book Review, *American Journal of Psychology*, Vol. 104, No.2, Summer 1991.

- Bonelli, Pierre, Parodi, Alexandre, Sen, Sandip and Wilson, Stewart, 'NEWBOOLE: A Fast GBML System', in Porter, Bruce and Mooney, Raymond, *Machine Learning: Proceedings of the Seventh International Conference*, Morgan Kaufmann, San Mateo, 1990.
- Bourbaki, Nick, 'Turing, Searle, & Thought', *AI EXPERT*, Vol. 5, No. 7, July, 1990, pps. 52-59.
- Bower, T.G.R., *A Primer of Infant Development*, W. H. Freeman & Company, San Francisco, 1977.
- Boyd, William C., 'Modern Ideas on Race, in the Light of Our Knowledge of Blood Groups and Other Characters with Known Mode of Inheritance', in Leone, Charles A. (Ed), *Taxonomic Biochemistry and Serology*, The Roland Press Company, New York, 1964.
- Brainerd, Charles J., *Piaget's Theory of Intelligence*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.
- Bratko, Ivan & Michie, Donald, 'Some comments on rule induction', in *The Knowledge Engineering Review*, Cambridge University Press, Cambridge, Vol. 2, No. 1, March 1987.
- Bree, D.S., & Smit, R., *Non Standard Uses of IF*, in Elithorn, Alick and Banerji, Ranan (Eds.), *Artificial and Human Intelligence*.
- Breiman, Leo, Friedman, Jerome H., Olshen, Richard A., and Stone, Charles J., *Classification and Regression Trees*, Wadsworth International Group, Belmont, California, 1984.
- Brillouin, Leon, *Science and Information Theory*, Academic Press, New York, 1956.
- Brooks, R.A., *Achieving artificial intelligence through building robots*, A.I. Memo 899, M.I.T. A.I. Lab., May 1986, (not seen, quoted in Cliff, D.T., *Computational Neuroethology; A Provisional Manifesto*, Cognitive Science Research Paper Serial No. CSRP 162, The University of Sussex, Brighton, May 1990).
- Brooks, Rodney A., *Intelligence Without Reason*, Proceedings of the Twelfth International Conference on Artificial Intelligence, Volume 2, August 1991, pps. 569 - 595.

Brown, Judith C., *Immodest Acts*, Oxford University Press, Oxford, 1986.

Brown, P.J., *Functions for selecting tests in diagnostic key construction*, *Biometrika*, Vol. 64, pp. 589 - 596, 1977; referenced by Dunn & Everitt, 1982.

Bruner, J. S., Goodnow, J. J., and Austin, G. A., *A Study of Thinking*, Wiley, New York, 1956; (not seen), referred to in Hunt *et al.*

Buford, Thomas O. (Ed.), *Essays on Other Minds*, University of Illinois Press, Urbana, U.S.A. 1970.

Bullock, T.H., "In search of principles in neural integration", in: Fentress, J. (Ed.), *Simpler Networks and behaviour*, Sinauer Assoc., Sunderland, Mass., 1976, pps. 52-60 (not seen, quoted in Miall, Christopher, "The Diversity of Neuronal Properties", in: Durbin, Richard, Miall, Christopher and Mitchison, Graeme (Eds.), *The Computing Neuron*, Addison-Wesley, England, 1989).

Buntine, Wray, 'Decision Tree Induction Systems: A Bayesian Analysis', in *Uncertainty in Artificial Analysis*, publisher unknown, Seattle, 10 July 1987.

Burr, Irving W., *Engineering Statistics and Quality Control*, McGraw-Hill Book Company, New York, 1953.

Butler, Eamonn and Pirie, Madsen, *Test Your IQ*, Pan Books, London, 1983.

Carlson, B. C., *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.

Carter, Chris and Catlett, Jason, 'Credit Assessment using Machine Learning', *IEEE Expert*, Fall 1987.

Catlett, J., 'Peephaling: choosing attributes efficiently for megainduction', in Sleeman, Derek and Edwards, Peter, *Machine Learning Proceedings of the Ninth International Workshop*, Morgan Kaufmann Incorporated, 1992.

Caudill, Maureen, *Neural Network Training Tips and Techniques*, *AI Expert*, January 1991.

Cestnik, Bojan, Kononenko, Igor and Bratko, Ivan, 'Assistant 86: A Knowledge-Elicitation Tool for Sophisticated Users', in Bratko, I. and Lavrac, N., *Progress in Machine Learning*, Sigma Press, England, 1987.

Chalmers, A.F., *What is this thing called Science*, Second Edition, University of Queensland Press, St Lucia, 1982.

Chaplin, J.P., *Dictionary of Psychology*, Dell Publishing Co., New York, 1975.

Cheng, Jie, Fayyad, Usama M., Irani, Keki B. and Qian, Zhaogang, 'Improved Decision Trees: A Generalised Version of ID3', in Laird, John (Ed.), *Proceedings of the Fifth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Mateo, U.S.A., 1988.

Churchland, Paul M. and Churchland, Patricia Smith, 'Could a Machine Think', *Scientific American*, Vol. 262 No. 1, January 1990.

Churchland, Paul M., *A Neurocomputational Perspective*, The MIT Press, Cambridge, Massachusetts, 1992.

Chwedorowicz, Józef, 'Origin, structure and function of fuzzy beliefs', in Zétényi, Tamás (Ed.), *Fuzzy Sets in Psychology*, North-Holland, Amsterdam, 1988.

Cliff, D.T., *Computational Neuroethology; A Provisional Manifesto*, Cognitive Science Research Paper Serial No. CSRP 162, The University of Sussex, Brighton, May 1990.

Coady, C. A. J., *Testimony*, Oxford University Press, Oxford, 1992.

Cohen, L. Jonathon, & Hesse, Mary, (Ed.), *Applications of Inductive Logic*, Clarendon Press, Oxford, 1980.

Cohen, Paul R., & Feigenbaum, Edward A., *The Handbook of Artificial Intelligence*, Vol.3, HeurisTech Press, Stanford, California, 1982.

Collier, P. A., *Manual for TL*, unpublished manuscript.

Collier, P.A. and Faulkner, E.G., *Decision Tree Generation using Statistical Methods & a comparison with other methods*, Second International Symposium on Artificial Intelligence, Monterrey, Mexico, 1989.

- Collier, P.A. and Faulkner, E.G. *Interactive Decision Tree Generation using Statistical Methods*, Australian Joint Artificial Intelligence Conference, Melbourne, 1989.
- Collier, P.A., 'Computer Key Generation from Quantitative Data', unpublished manuscript 1988.
- Collier, P.A., *Inductive Inference for Botanical Keys*, in Proceedings of the Third Australian Conference on Applications of Expert Systems, The New South Wales Institute of Technology, Sydney, 1987.
- Collier, P.A., *Inductive Inference for Botanical Keys*, R87-1, Information Science Department, University of Tasmania, Hobart, 1987.
- Collins, H.M., 'Domains in Which Expert Systems Could Succeed', *Third International Expert Systems Conference*, Learned Information Inc., Oxford, 1987.
- Copi, Irving M., *Introduction to logic*, fifth edition, Macmillan Publishing Company, New York, 1978.
- Coyne, Anthony M., *Introduction to Inductive Reasoning*, University Press of America, Inc., London, 1984.
- Crick, Francis and Koch, Christof, 'The Problem of Consciousness', *Scientific American*, Vol. 267 No. 3, September 1992.
- Crick, Francis and Koch, Christof, *Towards a Neurobiological Theory of Consciousness*, CNS Memo 9, January 28, 1991.
- Cromer, Alan, *Uncommon Sense*, Oxford University Press, Oxford, 1993.
- Cronbach, Lee J., 'On the non-rational application of information measures in psychology', in Quastler, Henry, (Ed.), *Information Theory in Psychology*, The Free Press, Glencoe, Illinois, 1955.
- Crosson, Frederick J., *Human and Artificial Intelligence*, Appleton-Century-Crofts, New York, 1970.
- Crowson, R.A., *Classification and Biology*, Heinemann Educational Books Ltd., London, 1970.

Czuchy, Andrew J., *A Neural Network Instantiation Environment*, Dr. Dobbs Journal, M&T Publishing Inc., Redwood City, California. April 1990.

Davidson, Clive, "Common sense & the computer", *New Scientist*, 2 April 1994, IPC Magazines Ltd., England, 1994.

*DARPA Neural Network Study*, AFCEA International Press, Fairfax, 1988 (not seen, quoted in Bach, Ivan N., *Data Complexity*, Neuron-Digest, Vol. 5, No. 51, December 1989).

Darwin, C., *On the Origin of Species*, Murray, London, 1859, (not seen); quoted in Cain, A. J., 'The Assessment of New Types of Character in Taxonomy', in Hawkes, J. G.(Ed.), *Chemotaxonomy and Serotaxonomy*, Academic Press, London, 1968.

Dawkins, Richard, *The Blind Watchmaker*, Longman Scientific & Technical, Harlow, England, 1986.

Dennett, Daniel C., *Elbow Room*, The MIT Press, Cambridge Massachusetts, 1984.

de Wit, H. C. D., *Plants of the World - The Higher Plants*, Volume 1, Thames and Hudson, London, 1963.

Dhillon, Balbir S., *Quality Control, Reliability, and Engineering Design*, Marcel Dekker, Inc., New York, 1985.

Dietterich, Thomas G., 'Limitations on Inductive Learning', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishers Inc., San Mateo, U.S.A., 1989.

Dietterich, Thomas G., Hild, Hermann and Bakiri, Ghulum, 'A Comparative study of ID3 and Backpropagation for English Text-to-Speech Mapping', in Porter, Bruce and Mooney, Raymond, *Machine Learning: Proceedings of the Seventh International Conference*, Morgan Kaufmann, San Mateo, 1990.

Ditlea, Steve, 'Artificial Intelligence', *Omni*, Volume 9, Number 7, Omni Publications International Ltd., New York, April 1987.



- Dreyfuss, G. R. *et al.*, 'On the psycholinguistic reality of fuzzy sets', in *Functionalism*, Grossman, R. *et al.* (Eds.), Chicago, IL: Univ. Chicago 1975, pps. 135-149 (not seen).
- Duda, Richard O., Hart, Peter E., and Nilsson, Nils J., *Subjective Bayesian methods for rule-based inference systems*, Proceedings of the National Computer Conference, AFIPS, 45, 1976.
- Dunn, G., and Everitt, B.S., *An introduction to mathematical taxonomy*, Cambridge University Press, Cambridge, 1982.
- Durbin, Richard, Miall, Christopher and Mitchison, Graeme (Eds.) *The Computing Neuron*, Addison-Wesley, Wokingham, England, 1989.
- Durkin, *AI Expert*, April 1992.
- Eccles, Sir John, *The Synapse*, Physiological Psychology, W. H. Freeman and Company, San Francisco, 1972.
- Edgington, Eugene S., *The Distribution-free approach*, McGraw-Hill, New York, 1969.
- Edelman, Gerald M., *The Remembered Present: A Biological Theory of Consciousness*, Basic Books Inc., New York, 1989.
- Elithorn, Alick and Banerji, Ranan (eds), *Artificial and Human Intelligence*, Elsevier Science Publications B.V., Amsterdam, 1984.
- English, Horace B., and English, Ava C., *A Comprehensive Dictionary of Psychological and Psychoanalytic Terms*, Longmans, Green and Co., New York, 1961.
- Erdtman, Holger, 'The Assessment of Biochemical Techniques in Plant Taxonomy', in Hawkes, J. G. (Ed.), *Chemotaxonomy and Serotaxonomy*, Academic Press, London, 1968.
- Evans, D. F., Hill, M. O. & Ward, S. D., *A dichotomous key to British submontane vegetation*, Occasional Paper No. 1, Institute for Terrestrial Ecology, Bangor, North Wales, 1977.
- Evans, Peter and Deehan, Geoff, *The Descent of Mind: The Purpose and Nature of Intelligence*, Grafton Books, London, 1990.
- Everitt, B.S., 'Unresolved Problems in Cluster Analysis', *Biometrics*, 35, pps. 169-181, 1979; (not seen), referenced by the SAS Institute

Inc, *SAS/STAT User's Guide, Release 6.03*, SAS Institute Inc., Cary, NC, 1988.

Everitt, B.S., *Cluster Analysis*, 2nd Edition, Heinemann Educational Books Ltd., London, 1980; (not seen), referenced by the SAS Institute Inc, *SAS/STAT User's Guide, Release 6.03*, SAS Institute Inc., Cary, NC, 1988.

Exton, Harold, *Multiple Hypergeometric Functions and Applications*, Ellis Horwood Limited, 1976.

Eysenck, H.J., *Know Your Own IQ*, Penguin Books, Harmondsworth, Middlesex, 1962.

Fahlman, Scott E., *An Empirical Study of Learning Speed in Back-Propagation Networks*, CMU-CS-88-162, Carnegie-Mellon Report, September 1988.

Fahlman. Scott E., "Faster-Learning Variations on Back-Propagation: An Empirical Study" in *Proceedings of the 1988 Connectionist Models Summer School*, Morgan Kaufmann, 1988, (not seen, referred to in Fahlman, Scott E. and Leblere, Christian, *The Cascade-Correlation Learning Architecture*, Report CMU-CS-90-100, Carnegie Mellon University, Pittsburgh, Feb. 1990.).

Fahlman. Scott E. and Leblere, Christian, *The Cascade-Correlation Learning Architecture*, Report CMU-CS-90-100, Carnegie Mellon University, Pittsburgh, Feb. 1990.

Ferguson, Andrew, *Biochemical Systematics and Evolution*, John Wiley and Sons, New York, 1980.

Ferry, Georgina, 'The Egalitarian Brain', *New Scientist*, Volume 109, Number 1490, 9 January 1986.

Fischbach, Gerald D., 'Mind and Brain', *Scientific American*, Vol. 267 No. 3, September 1992.

Fisher, R.A. 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics*, 7, pps. 179-188, 1936; (not seen) referenced in a public communication by nlonginow@falcon.aamrl.wpafb.af.mil.

Forsyth, R. S., 'The Evolution of Intelligence', in *Third International Expert Systems Conference*, Learned Information Ltd. (Ed.), London, 1987.

Freeling, Anthony N. S., 'Fuzzy Sets and Decision Analysis', *IEEE Transactions on Systems, Man, and Cybernetics*, Volume SMC-10, Number 7, July 1980.

Fu, Li-Min, 'Learning Object-Level and Meta-Level Knowledge in Expert Systems', Technical Report No. STAN-CS-86-1091, Department of Computer Science, Stanford University, 1985.

Fukushima, Kuniyoko, Miyake, Sei and Ito, Takayuki, *Neocognition: A Neural Network Model for a Mechanism of Visual Pattern Recognition*, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. SMC-13, No. 5, September/October 1983.

Gabora, Liane and Collins, Rob (Eds.), *Alife Digest*, Artificial Life Research Group, UCLA, Los Angeles, Volume #088, October 28th, 1992.

Galton, Francis, *Hereditary Genius*, Collins, London, U.K. 1962; (a reprint of the text and diagrams of Galton's second edition of *Hereditary Genius*, published by Macmillan & Co. Ltd., 1869).

Gardner, Lytt I., 'Deprivation in Dwarfism', in *The Nature and Nurture of Behaviour, Readings from the Scientific American*, W. H. Freeman and Company, San Francisco, 1973.

Garey, M.R., *Optimal binary identification procedures*, *SIAM Journal of Applied Mathematics*, Vol. 23, pps. 173 - 186, 1972; (not seen) referenced in Dunn & Everitt.

Garey, M.R. & Graham, R.L., *Performance bounds on the splitting algorithm for binary testing*, *Acta Informatica*, Vol. 3, pps. 347 - 355, 1974; (not seen), referenced in Dunn & Everitt.

Garner and McGill, 'The relation between information and variance analysis', *Psychometrika*, Volume 21, 1956 (not seen), pps. 219-228, referred to in: Macnaughton-Smith, P., *Some Statistical and Other Numerical Techniques for Classifying Individuals*, Her Majesty's Stationery Office, London, 1965.

- Garner, W. R. and McGill, William J., 'The relation between information and variance analysis', *Psychometrika*, Volume 21, No. 3, September 1956.
- Garrett, Henry E., and Woodworth., R.S., *Statistics in Psychology and Education*, Vakils, Feffer and Simons Pty. Ltd., Bombay, 1967.
- Gaschnig, J., 'Prospector: An expert system for mineral exploration', in Michie, Donald, *Introductory Readings in EXPERT SYSTEMS*.
- Gevarter, William B., 'The Nature and Evaluation of Commercial Expert System Building Tools', *Computer*, Volume 20, Number 5, I.E.E.E., New York, May 1987.
- Goldman-Rakic, Patricia S., 'Working Memory and the Mind', *Scientific American*, Vol. 267 No. 3, September 1992.
- Goodall, D. W., 'Objective Methods for the Classification of Vegetation', *Australian Journal of Botany*, Volume 2, Number 1, Commonwealth Scientific and Industrial Research Organisation, East Melbourne, February 1954.
- Gower, J. C., 'Relating Classification to Identification', in Pankhurst, R. J., (Ed.), *Biological Identification with Computers*,. Systematics Association Special Volume No. 7, Academic Press, London, 1975.
- Gower, J. C. and Barnett, J. A., 'Selecting Tests in Diagnostic Keys with Unknown Responses', *Nature*, Vol. 232, August 13<sup>th</sup> 1971.
- Gower, J.C., and Payne, R.W., 'A comparison of different criteria for selecting binary tests in diagnostic keys', in *Biometrika*, Vol. 62 No. 3, 1975.
- Grefenstette, John J. and Ramsey, Connie Loggia, 'An Approach to Anytime Learning', in Sleeman, Derek and Edwards, Peter, *Machine Learning Proceedings of the Ninth International Workshop*, Morgan Kaufmann Incorporated, 1992.
- Gribbin, John, "Is the Universe alive?", *New Scientist*, New Science Publications, London, 15 January, 1994.
- Groves, Phillip and Schlesinger, Kurt, *Biological Psychology*, Wm. C. Brown Company, Dubuque, Iowa, 1979.

- Gryna, Frank M., *Basic Statistical Methods*, in Juran, J. M., Gryna, Jr., Dr. Frank M., Bingham, Jr., R. S., (Eds.), *Quality Control Handbook*.
- Hadingham, Paul T., *Towards a neural net architecture for rapid learning in machine vision*, Proceedings of the SPIE Conference on Automatic Inspection and High Speed Vision Architectures III, Philadelphia, Pennsylvania, 5-10 November, 1989. (This may also be obtained as Technical Report 89/16, Department of Computer Science, University of Western Australia).
- Hamming, Richard W., *Coding and Information Theory*, Prentice-Hall Inc., New Jersey, 1980.
- Hamming, R. W., 'Error Detecting and Error Correcting Codes', in *Bell System Technical Journal*, Volume 26, Number 2, April, 1950, pps. 147-160.
- Hanson, Stephen Jose and Olson, Carl R., *Connectionist Modelling and Brain Function: The Developing Interface*, The MIT Press, Cambridge, Massachusetts, 1990.
- Harel, David, *Algorithmics The Spirit of Computing*, Addison-Wesley Publishing Company, Wokingham, England, 1987.
- Harlow, Harry F. and Harlow, Margaret Kuenne, 'Learning to Think', in *Physiological Psychology, Readings from the Scientific American*, W. H. Freeman and Company, San Francisco, 1972.
- Hart, Anna, *Knowledge Acquisition for Expert Systems*, Kogan Page, London, 1986.
- Harth, Erich, *Order and Chaos in Neural Systems: An Approach to the Dynamics of Higher Brain Functions*, IEEE Transactions on Systems, Man and Cybernetics, Vol SMC-13, No. 5, September/October, 1983.
- Hatcher, William S., *The Logical Foundations of Mathematics*, Pergamon Press, Oxford, 1982.
- Hayes-Roth, Frederick, Waterman, Donald A., and Lenat, Douglas B., (Eds.), *Building Expert Systems*, Addison-Wesley Publishing Company, Inc., London, 1983.
- Heathcote, Adrian, "False prophets muddy 'truth'", *The Australian*, 9 March 1994, Nationwide News Pty. Ltd., Canberra, 1994.

- Hebb, D.O., *The Organisation of Behaviour: A Neuropsychological Theory*, Wiley, New York, 1949, (not seen, referred to in Edelman, Gerald M., *The Remembered Present: A Biological Theory of Consciousness*, Basic Books Inc., New York, 1989, p. 38).
- Hecht-Nielsen, Robert, *Neurocomputing: picking the human brain*, IEEE Spectrum 25(3), March 1988.
- Helmholtz, H. Von, *Preliminary report on the velocity of the nerve impulse*, 1850, reprinted and translated in *Founders of Experimental Psychology*, Blasius, W., Boylan, J., and Kramer, K., Eds., Munich: 25<sup>th</sup> International Congress of Physiological Science, 1971.
- Hempel, Carl C., 'Studies in the Logic of Confirmation', in Luckenbach, Sidney A., (Ed.), *Probabilities, Problems and Paradoxes*,.
- Hennig, W., *Phylogenetic systematics*, University of Illinois Press, Urbana, U.S.A., 1966 (not seen), referred to in Humphries, Christopher J. and Parenti, Lynne R., *Cladistic Biogeography*, Clarendon Press, Oxford, 1989.
- Hillis, W. Daniel, 'The Connection Machine', in *Scientific American*, Scientific American Incorporated, New York, June 1987, Vol. 256, No. 6.
- Hinton, Geoffrey E., 'How Neural Networks Learn from Experience', *Scientific American*, Vol. 267 No. 3, September 1992.
- Hilts, Philip J., 'The Dean of Artificial Intelligence', *Psychology Today*, Volume 17, number 1, January 1983.
- Hoel, Paul. S., *Introduction to Mathematical Statistics*, John Wiley & Sons, New York, 1954.
- Hofstadter, Douglas R., *GÖDEL, ESCHER, BACH: An Eternal Golden Braid*, Penguin Books, Harmondsworth, England, 1982.
- Hofstadter, Douglas R. and Dennett, Daniel C., (Eds.), *THE MIND'S I*, Penguin Books, Harmondsworth, England, 1982.
- Holland, John H., Holyoak, Keith J., Nisbett, Richard E., Thagard, Paul R., *Induction*, The MIT Press, Cambridge, Massachusetts, 1987.

- Holloway, Marguerite, 'Rx for addiction', *Scientific American*, Volume 264, Number 3, March 1991.
- Horgan, John, 'Life in a Test Tube?', *Scientific American*, Vol. 266 No. 5, May 1992.
- Howell, Jim, 'S-Trees: A New Way to Handle Subjective Rules', *AI Expert*, Vol. 7, No. 2, February 1992.
- Hoyle, G., "Neural mechanisms underlying behaviour of invertebrates", in Gazzaniga, M.S. and Blakemore, C. (Eds.), *Handbook of Psychobiology*, pages 3-48, Academic Press, New York, 1975, (not seen, quoted in Cliff, D.T., *Computational Neuroethology; A Provisional Manifesto*, Cognitive Science Research Paper Serial No. CSRP 162, The University of Sussex, Brighton, May 1990).
- Humphrey, Nicholas, "The private world of consciousness", *New Scientist*, 8 January 1994, New Science Publications, London, 1994.
- Humphries, Christopher J. and Parenti, Lynne R., *Cladistic Biogeography*, Clarendon Press, Oxford, 1989.
- Hunt, E.B., Marin, J. and Stone, P. T., *Experiments in Induction*, Academic Press, New York, 1966.
- Huxley, Aldous, 'The Doors of Perception', in *The Doors of Perception and Heaven and Hell*, Penguin Books, Harmondsworth, England, 1961.
- Inhelder, Barbel, and Piaget, Jean, *The Growth of Logical Thinking from childhood to adolescence*, Routledge & Kegan Paul, London, 1958.
- Jain, Anil K. and Dubes, Richard C., *Algorithms for Clustering Data*, Prentice-Hall, New Jersey, 1988.
- Jelen, F. C., (Ed.), *Cost and Optimization Engineering*, McGraw-Hill Book Company, New York, 1970.
- Juran, J. M., Gryna, Jr., Dr. Frank M., Bingham, Jr., R. S., (Eds.), *Quality Control Handbook*, McGraw-Hill Book Company, New York, 1951.

- Kandel, Abraham and Byatt, William J., 'Fuzzy Sets, Fuzzy Algebra, and Fuzzy Statistics', *Proceedings of the I.E.E.E.*, Volume 66, No. 12, December 1978
- Kandel, Eric R. and Hawkins, Robert D., 'The Biological Basis of Learning and Individuality', *Scientific American*, Vol. 267 No. 3, September 1992.
- Keppel, Geoffrey, *Design & Analysis*, Prentice-Hall Inc., New Jersey, 1973.
- Keynes, Richard D., *The Nerve Impulses and the Squid*, Physiological Psychology, W.H.Freeman and Company, San Francisco, 1972.
- Kidd, A.L., 'Human factors in expert systems', in Coombes, K., (Ed.), *Proceedings of the Ergonomic Society Conference 1983*, Taylor and Francis, London, 1983, (not seen); quoted by Gammack, J.G., 'Modelling expert knowledge using cognitively compatible structures', in *Third International Expert Systems Conference*, Learned Information (Europe) Ltd, London, 1987.
- Kitzinger, Celia, 'Margaret Boden: Probing the mystery of the human mind', *The Psychologist*, Vol. 4, No. 1, January 1991, pps. 13-15.
- Klemke, E.D., Kline, A.David, and Hollinger, Robert (Eds.), *Philosophy The Basic Issues*, St. Martin's Press, New York, 1982.
- Klimasauskas, Casy, *Neural Nets and Noise Filtering*, Dr. Dobbs Journal of Software Tools, January 1989.
- Knowler, LLOYD A., Howell, John M., Gold, Ben K., Coleman, Edward P., Moan, Obert B., Knowler, William C., *Quality Control by Statistical Methods*, McGraw-Hill Book Company, New York, 1969.
- Koch, C., and Segev, I. (Eds.), *Methods in Neural Modelling: From Synapses to Networks*, Bradford Books, MIT Press, Cambridge, USA, 1989.
- Kohler, Heinz, *Statistics for Business and Economics*, Scott, Foresman and Company, London, 1985.
- Kohonen, Teuvo, *Associative Memory*, Springer-Verlag, Berlin, 1977.



Koppel, Tom, 'Profile: Supercomputer Solo', *Scientific American*, Volume 264, Number 3, March 1991.

Kowalski, Gary, *The Souls of Animals*, Stillpoint Publishing, Walpole, U.S.A., 1991.

Kreyszig, Erwin, *Advanced Engineering Mathematics*, Fifth Edition, John Wiley and Sons, New York, 1983.

Kroy, Moshe, *Moral Competence, An Application of Modal Logic to Rationalistic Psychology*, Mouton, The Hague, 1975,

Kullback, Solomon, *Information Theory and Statistics*, John Wiley & Sons, New York, 1959.

Kuncicky, David and Kandel, Abraham, 'The weighted fuzzy expected value as an activation function for the parallel distributed processing models', in Zétényi, Tamás (Ed.), *Fuzzy Sets in Psychology*, North-Holland, Amsterdam, 1988.

Labinowicz, Ed, *The Piaget Primer, Thinking, Learning, Teaching*, Addison-Wesley Publishing Company, Menlo Park, California, 1980.

Lawler, R.W., *Computer Experience and Cognitive Development*, Ellis Horwood Limited, West Sussex, England, 1985.

Lawrence, Jeanette, *Untangling Neural Nets*, Dr. Dobbs Journal, M&T Publishing Inc., Redwood City, California. April 1990.

Leibovic, K. Nicholas, *Phototransduction in Vertebrate Rods: An Example of the Interaction of Theory and Experiment in Neuroscience*; IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-13, No. 5, September/October, 1983.

Leighton, R., and Wieland, A., *The Aspirin/MIGRAINES Software Tools User's Manual, Release 4.0*, The MITRE Corporation, Washington, 1991.

Leitch, Roy & Francis, John, 'Towards Intelligent Control Systems', in Mamdani, Abe, & Efstathiou, Janet, (Eds.), *Expert Systems and Optimisation in Process Control*,.

Lewin, Roger, "I buzz therefore I think", *New Scientist*, 15 January 1994, New Science Publications, London, 1994.

- Lewin, Roger, "Scenes from a biological revolution", *New Scientist*, 5 March 1994, New Science Publications, London, 1994.
- Lewis, Edwin R., *The Elements of Single Neurons: A Review*, IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-13, No. 5, September/October, 1983.
- Linsker, R., "Self-Organisation in a Perceptual System: How Network Models and Information Theory May Shed Light on Neural Organisation", in: Hanson, Stephen Jose and Olson, Carl R., *Connectionist Modelling and Brain Function: The Developing Interface*, The MIT Press, Cambridge, Massachusetts, 1990.
- Lippmann, Richard L., *An Introduction to Computing with Neural Nets*, IEEE ASSP Magazine, April 1987.
- Long, Debra L., Graesser, Arthur C., Long, Charles J., *Four Computational Models for Investigating Neuropsychological Decision-making*, in: *Cognitive Approaches to Neuropsychology*, Williams, J. Michael, and Long, Charles J., Eds., Plenum Press, New York, 1988.
- Lovelock, James, *Healing Gaia Practical Medicine for the Planet*, Harmony Books, New York, 1991.
- Luce, A. A., *Teach Yourself Logic*, English Universities Press, London, 1958.
- Luckenbach, Sidney A., (Ed.), *Probabilities, Problems and Paradoxes*, Dickenson Publishing Company Inc., Encino, California, 1972.
- Lycan, William G. (Ed.), *Mind and Cognition*, Blackwell, 1990.
- MACazine Eds, 'DATELINE:Macintosh', in *MACazine*, Icon Concepts Corporation, Austin Texas, October 1987.
- Mackie, J. L., 'The Paradox of Confirmation', in Luckenbach, Sidney A., (Ed.), *Probabilities, Problems and Paradoxes*.
- Macnaughton-Smith, P., *Some Statistical and Other Numerical Techniques for Classifying Individuals*, Her Majesty's Stationery Office, London, 1965.
- MacWilliams, F.J. and Sloane, N.J.A., *The Theory of Error-Correcting Codes*, North-Holland Publishing Company, Amsterdam, 1978.

Mamdani, Abe, & Efstathiou, Janet, (Eds.), *Expert Systems and Optimisation in Process Control*, Gower Technical Press, Aldershot, England, 1986.

Mao, Chengjiang, 'THOUGHT: An Integrated Learning system for Acquiring Knowledge Structure', in Sleeman, Derek and Edwards, Peter, *Machine Learning Proceedings of the Ninth International Workshop*, Morgan Kaufmann Incorporated, 1992.

Marshall, John C., *Sensation and Semantics*, Nature, Vol. 334 4, Aug 1988.

Matheus, Christopher, 'A Constructive Induction Framework', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishing Inc., San Mateo, U.S.A., 1989.

Mayr, E, (1959) quoted in Sokal, Robert R. and Sneath, Peter H. A., *Principles of Numerical Taxonomy*, W. H. Freeman and Company, San Francisco, 1963.

Mel, Bartlett W., *The Sigma-Pi Column: A Model of Associative Learning in Cerebral Neocortex*, CNS Memo 6, California Institute of Technology, California, 30<sup>th</sup> April 1990.

Mayes, J. Terry, Draper, Stephen W., McGregor, Alison M. and Oatley, Keith, "Information Flow in a User Interface: The Effect of Experience and Context on the Recall of MacWrite Screens", in Preece, Jenny and Keller, Laurie (Eds.), *Human-Computer Interaction*, Prentice Hall, Hertfordshire, England, 1989.

Margalef, D. Ramon, 'Information Theory in Ecology', in *General Systems*, Volume 3, p. 36, 1958. p. 68. This paper was originally in Spanish and was presented by the author to the Royal Academy of Sciences and Arts of Barcelona on the occasion of his acceptance of election to the Academy on April 4, 1957. The English-language version was translated by Wendell Hall from *Memorias de la Real Academia de Ciencias y Artes de Barcelona*, 23: 373-449, November, 1957.

McCallum, R. Andrew and Spackman, Kent A., 'Using Genetic Algorithms to Learn Disjunctive Rules from Examples', in Porter, Bruce and Mooney, Raymond, *Machine Learning: Proceedings of the*

*Seventh International Conference*, Morgan Kaufmann, San Mateo, 1990.

McCarthy, Rosaleen and Warrington, Elizabeth K., *Cognitive Neuropsychology A Clinical Introduction*, Academic Press, San Diego, 1990.

McCulloch, W.S., and Pitts, W., *A Logical Calculus of the Ideas Imminent in Nervous Activity*, Bulletin of Mathematical Biophysics, 5, 1943.

McGill, William and Quastler, Henry, 'Standardised Nomenclature: An Attempt', in Quastler, Henry (Ed.), Quastler, Henry, (Ed.), *Information Theory in Psychology*, The Free Press, Glencoe, Illinois, 1955.

McLaren, Ian, "The Computational Unit as an Assembly of Neurones: an Implementation of an Error Correcting Learning Algorithm", in: Durbin, Richard, Miall, Christopher and Mitchison, Graeme (Eds.), *The Computing Neuron*, Addison-Wesley, England, 1989, pps. 160 - 179.

McPherson, D.G., University of Tasmania, Hobart, untitled and undated lecture handout.

Massaro, Dominic W., Book Review, *American Journal of Psychology*, Vol. 104, No.2, Summer 1991.

Michalski, Ryszard S., Carbonell, Jaime G., & Mitchell, Tom M., (Eds.), *Machine Learning, An Artificial Intelligence Approach*, Tioga Publishing Company, Palo Alto, 1983.

Michalski, Ryszard S., Chilausky, R.L., 'Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis', in *International Journal of Policy Analysis and Information Systems*, Vol. 4, No. 2, June 1980.

Michie, D., 'Current developments in Expert systems', *Proceedings of the Second Australian Conference on Applications of Expert Systems*, Sydney, 1986.

Michie, Donald, *Introductory Readings in EXPERT SYSTEMS*, Gordon and Breach Science Publishers, New York, 1982.

Mill, John Stuart, *A System of Logic Ratiocinative and Inductive*, eighth edition, Longmans, Green and Co., London, 1884.

Miller, George A., *Psychology, The Science of Mental Life*, Penguin Books, Harmondsworth, England, 1972.

Milligan, G.W., "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms", *Psychometrika*, 45, 1980, pps. 325-342, (not seen), referenced in SAS Institute Inc, *SAS/STAT User's Guide, Release 6.03*, SAS Institute Inc., Cary, NC, 1988.

Minsky, Marvin L., quoted in Boden, Margaret A., *Artificial Intelligence and Natural Man*.

Minsky, Marvin L., *The Society of Mind*, Simon and Schuster, New York, 1986.

Minsky, Marvin L., and Papert, Seymour A., *Perceptrons*, The MIT Press, Cambridge, Massachusetts, 1988.

Mishkin, Mortimer & Appenzeller, Tim, *The Anatomy of Memory*: Scientific American, Scientific American Inc., June 1987, Vol. 256, No.6.

Modesitt, K.L., 'Experts: Human and Otherwise', in *Proceedings of the Third International Expert Systems Conference*, Learned Information Ltd., Oxford, 1987.

Møller, Martin F., *A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning*, Preprint PB-339, Computer Science Department, University of Aarhus, Denmark, November 13, 1990.

Moroney, M. J., *Facts from Figures*, Penguin Books Ltd., Harmondsworth, England, 1984.

Morse, L.E., *Specimen identification and key construction with time sharing computers*, Taxon, Vol. 20, pps. 269 - 282, 1971; (not seen), referenced in Dunn & Everitt.

Mouradian, William H., 'Knowledge Acquisition in a Medical Domain', *AI EXPERT*, Vol. 5, No. 7, July 1990, pps. 35-38.

Muggleton, Stephen, *Inductive Acquisition of Expert Knowledge*, Addison-Wesley, England, 1990.

Nash, John, *Developmental Psychology*, Prentice/Hall International Inc., London, England, 1973.

Negroponte, Nicholas, *Soft Architecture Machines*, Cambridge, MA., the MIT Press, 1975, (not seen), reference for quotation in Baecker, R.M., Buxton, W.A.S., "An Historical and Intellectual Perspective", in Preece, Jenny and Keller, Laurie (Eds.), *Human-Computer Interaction*, Prentice Hall, Hertfordshire, Great Britain, 1990.

Nelson, Gareth and Platnick, Norman, *Systematics and Biogeography, Cladistics and Vicariance*, Columbia University Press, New York, 1981.

Newquist III, Harvey P., "The Other Side of AI", in *AI EXPERT*, Volume 7, No. 3, March 1992.

New Scientist, 'New Scientist does not believe in fairies, flying saucers or Artificial Intelligence', *New Scientist*, 8 November 1984, IPC Magazines Ltd., England, 1984.

O'Neill, J.L., 'Plausible Reasoning', in *The Australian Computer Journal*, Vol. 19, No. 1, February 1987.

Orchard, A. E., 'Revision of the *Acaena Ovina* A. Cunn. (Rosaceae) Complex in Australia', *Trans. Roy. Soc. S. Aust.* (1969), Vol. 93.

Orlóci, László, 'Information Analysis in Phytosociology: Partition, Classification and Prediction', *Journal of Theoretical Biology*, Volume 20, 1968, pps. 271-284.

Orlóci, László, 'Information Analysis of Structure in Biological Collections', *Nature*, Volume 223, August 2, 1969.

Orlóci, László, 'Information theory models for hierarchical and non-hierarchical classifications', in Cole, A. J. (Ed.), *Numerical Taxonomy*, Proceedings of the Colloquium in Numerical Taxonomy Held in the University of St. Andrews, September 1968, pps. 148-164, Academic Press, London.

Orlóci, László, *Multivariate Analysis in Vegetation Research*, Dr. W. Junk, The Hague, 1978.

- Pankhurst, Richard J., 'A computer program for generating diagnostic keys', *Computer Journal* Vol. 13 No. 2, May 1970a.
- Pankhurst, Richard J., 'A computer program for generating diagnostic keys', *New Phytologist*, Vol. 62, pps. 35 - 43, 1970; (not seen), referenced in Dunn & Everitt. However Pankhurst comments 'The reference ... is completely fictitious! Presumably the Computer Journal paper is intended' , (private communication). The Computer Journal reference is listed above.
- Pankhurst, Richard J., 'An interactive program for the construction of identification keys', *Taxon*, August 1988.
- Pankhurst, R. J., (Ed.), *Biological Identification with Computers*, Systematics Association Special Volume No. 7, Academic Press, London, 1975.
- Pankhurst, Richard J., 'Botanical Keys Generated by Computer', *Watsonia*, 8 1971.
- Pankhurst, Richard J., 'Key generation by Computer', *Nature*, London, Vol. 227, September 19, 1970b.
- Pankhurst, Richard J., *Practical taxonomic computing*, Cambridge University Press, Cambridge, 1991.
- Pao, Yoh-Han, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1989.
- Partridge, D., 'Is Intuitive Expertise Rule Based?', in *Proceedings of the Third International Expert Systems Conference*, Learned Information Ltd., Oxford, 1987.
- Payne, R. W. and Preece, D. A., 'Identification keys and diagnostic tables: a review', *Journal of the Royal Statistical Society, Series A*, Volume 143, 1980, pps. 253-292.
- Payne, R. W., 'Genkey: a program for constructing diagnostic keys', in Pankhurst, R. J., (Ed.), *Biological Identification with Computers*, pps. 65 - 72.
- Pearson, Karl, *Contributions to the mathematical theory of evolution. I. Dissection of frequency curves*, Phil. Trans. R. Soc., A 185, 1894, (not seen) referenced in Gower, J. C., 'Relating Classification to

Identification', in Pankhurst, R. J. (Ed.), *Biological Identification with Computers*.

Pellegrino, James W., 'Inductive Reasoning Ability', in Sternberg, Robert J., (Ed.), *Human Abilities*, W.H. Freeman & Company, New York, 1985.

Peterson, W. Wesley and Weldon Jr., E.J., *Error Correcting Codes*, The MIT Press, Cambridge, Massachusetts, 1972.

Piaget, Jean, *The Child's Conception of Physical Causality*, Kegan Paul, Trench, Trubner & Co. Ltd., London, 1930.

Piaget, Jean, *The Origin of Intelligence in the Child*, Penguin Books, Harmondsworth, England, 1983.

Pielou, E. C., *The Measurement of Diversity in Different Types of Biological Collections*, Journal of Theoretical Biology, Volume 13, 1966, pps. 131-144.

Pine, Milton, 'Western Philosophy and Expert Systems', *Professional Computing*, Peter Isaacson Publications, Victoria, Australia, October 1989.

Popham, W. James, and Sirotnik, Kenneth A., *Educational Statistics, Use and Interpretation, 2nd Edition*, Harper & Row, New York, 1973.

Preece, Jenny and Keller, Laurie (Eds.), *Human-Computer Interaction*, Prentice Hall, Hertfordshire, Great Britain, 1990.

Putnam, H., 'Robots: Machines or artificially created life', in Crosson, Frederick J., (Ed.), *Human and Artificial Intelligence*.

Quastler, Henry, (Ed.), *Information Theory in Psychology*, The Free Press, Glencoe, Illinois, 1955.

Quastler, H., 'The measure of specificity' (not seen), in Quastler, H (Ed.), *Information Theory in Biology*, University of Illinois Press, Urbana, 1953 (not seen), referenced by McGill, William J., 'Isomorphism in Statistical Analysis', in Quastler, Henry (Ed.), *Information Theory in Psychology*, The Free Press, Glencoe, Illinois, 1953.



- Quastler, R., 'The measure of specificity', in Quastler, R (Ed.), *Information Theory in Biology*, University of Illinois Press, Urbana, 1953 (not seen), referenced in Margalef, D. R., *General Systems*, Volume 3, p. 36, 1958. However the initial attributed to Margalef's useage by his translator is probably in error, as William J. McGill references a paper of the same name in a volume of identical title published in the same year as being by H. Quastler, see: Quastler, H., 'The measure of specificity' (not seen), in Quastler, H (Ed.), *Information Theory in Biology*, University of Illinois Press, Urbana, 1953 (not seen), referenced by McGill, William J., 'Isomorphism in Statistical Analysis', in Quastler, Henry (Ed.), *Information Theory in Psychology*, The Free Press, Glencoe, Illinois, 1953.
- Quinlan, J. Ross, 'Decision Trees as Probabilistic Classifiers', in Langley, Pat (Ed.), *Proceedings of the Fourth International Workshop on Machine Learning*, June 1987, Morgan Kaufmann Publishers, Inc., San Mateo, U.S.A., 1987.
- Quinlan, J. Ross, 'Learning Efficient Classification Procedures and their Application to Chess End Games', in Michalski, Ryszard S., Carbonell, Jaime G., & Mitchell, Tom M., (Eds.), *Machine Learning, An Artificial Intelligence Approach*.
- Quinlan, J. Ross, *Induction of Decision Trees*, Technical Report 85.6, School of Computing Sciences, New South Wales Institute of Technology, 1985.
- Quinlan, J. Ross, 'Induction of Decision Trees', in *Machine Learning*, Vol. 1, No. 1.
- Quinlan, J. Ross, 'Learning with Continuous Classes', in Adams, Anthony and Sterling, Leon (Eds.), *Proceedings of the 5<sup>th</sup> Australian Joint Conference on Artificial Intelligence*, World Scientific Publishing Co., Singapore, November 1992.
- Quinlan, J. Ross, *Simplifying Decision Trees*, Technical Report 87.4, New South Wales Institute of Technology, Sydney, 1987.
- Quinlan, Ross J., 'Unknown Attribute Values in Induction', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishing Inc., San Mateo, U.S.A., 1989.

- Reason, J. T., 'Motion sickness, some theoretical considerations', *Int. J. of Man-Mach. Studies*, Volume 1, pps. 21-38, 1969 (not seen).
- Restak, Richard M., *The Brain*, Bantam Books, Toronto, 1984.
- Rendell, Larry, 'Comparing Systems and Analysing Functions to Improve Constructive Induction', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishing Inc., San Mateo, U.S.A., 1989.
- Rich, Elaine, *Artificial Intelligence*, McGraw-Hill Book Company, Singapore, 1983.
- Ridley, Mark, *Evolution and Classification, The Reformation of Cladism*, Longman, London, 1986.
- Ripley, B. D., 'Statistical Aspects of Neural Networks', invited lecture for SemStat (Séminaire Européen de Statistique), Sandbjerg, Denmark, 25-30 April 1992. To appear in the proceedings to be published by Chapman & Hall in January 1993,
- Rolls, Edmund, "The Representation and Storage of Information in Neuronal Networks in the Primate Cerebral Cortex and Hippocampus" in: Durbin, Richard, Miall, Christopher and Mitchison, Graeme (Eds.), *The Computing Neuron*, Addison-Wesley, England, 1989, pps. 125 - 159.
- Rosenblatt, Frank, *Principles of Neurodynamics: Perceptrons and the theory of Brain Mechanisms*, Spartan books, Washington, D.C., 1961.
- Rowe, Helga A. H., *Problem Solving and Intelligence*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1985.
- Rumelhart, David E. and McClelland, James L., *Parallel Distributed Processing*, MIT Press, Cambridge, Massachusetts, 1986.
- Russell, Bertrand, 'Analogy', in Buford, Thomas O., (Ed.) *Essays on Other Minds*.
- Sacks, Oliver, *The Man Who Mistook His Wife For His Hat*, Pan Books, London, 1986.
- SAS Institute Inc, *SAS/STAT User's Guide, Release 6.03*, SAS Institute Inc., Cary, NC, 1988.

Saxena, Sharad, 'Evaluating Alternative Instance Representation', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishing Inc., San Mateo, U.S.A., 1989.

Schank, Roger, *A.I. Magazine*, Winter/spring 1983, quoted by Amoliar, Stephen W., *Induction: Processes of Inference, Learning and Discovery*, IEEE Expert, Computer Society of the IEEE, USA, Fall 1987.

Scriven, Michael, 'The compleat robot; A prolegomena to androidology', in Crosson, Frederick J., (Ed.), *Human and Artificial Intelligence*.

Searle, John R. 'Consciousness, explanatory inversion, and cognitive science', *Behavioural and Brain Sciences*, 13, 1990.

Searle, John R., *Is the Brain's Mind a Computer Program?*, *Scientific American*, Vol. 262, No.1, January 1990.

Searle, John R., 'Minds, Brains and Programs', *The Behavioural and Brain Sciences* 3, 1980.

Sechu, Sundaram, 'On an Improved Diagnosis Program, *I.E.E.E. Transactions on Electronic Computers*; February 1965, Vol. EC-14.

Seshu, S., and Freeman, D. N., 'The diagnosis of asynchronous sequential switching systems', *IRE Trans. on Electronic Computers*, Vol. EC-11, August 1962.

Sejnowski, T. and Churchland, P., 'Silicon Brains' in *Australian Personal Computer*, Vol. 13 No. 11, Computer Publications Pty. Ltd., November 1992.

Sellars, Wilfred, *Are there non-deductive logics?*, in Luckenbach, Sidney A., *Probabilities, Problems, and Paradoxes*.

Sestito, Sabrina and Dillon, Tharam, *Using neural networks for the extraction of high level knowledge representations for machine learning*, Technical Report No. 5/89, May, 1989, Department of Computer Science, LaTrobe University, Victoria, Australia, 3083.

Shannon, Claude E., *A mathematical theory of communication*, *Bell System Technical Journal*, Vol. 27, pps. 379 - 423, 623-656, 1948.

- Shannon, Claude E. and Weaver, Warren, *The Mathematical Theory of Communication*, 12<sup>th</sup> Edition, University of Illinois Press, Urbana, U.S.A., September 1949.
- Sharkey, Noel E., *Neural Network Learning Techniques*, in McTear, Michael (Ed.), *Understanding Cognitive Science*, Ellis Horwood Limited, Chichester, 1988.
- Shatz, Carla J., 'The Developing Brain', *Scientific American*, Vol. 267 No. 3, September 1992.
- Shwayder, K., *Conversion of limited entry decision tables to computer programs - a proposed modification to Pollack's algorithm*, *Communications of the Association of Computing Machinery*, Vol. 14, pps. 69 - 73, 1971; referenced in Dunn & Everitt.
- Shwayder, K., *Extending the information theory approach to converting limited-entry decision tables to computer programs*, *Communications of the Association of Computing Machinery*, Vol. 17, pps. 532 - 37, 1974; referenced in Dunn & Everitt.
- Siegler, R. S. & Kotovsky, K., 'Two levels of giftedness' in Sternberg, Robert J. & Davidson, Janet E., (Eds.), *Conceptions of Giftedness*.
- Simon, Herbert A., *The Sciences of the Artificial*, Second Edition, The MIT Press, Cambridge, Massachusetts, 1985.
- Sloman, Towards *a computational theory of mind*, in Yazdani, Masoud and Narayanan, Ajit (Eds.), *Artificial Intelligence: human effects*.
- Smart, Mollie S., & Smart, Russell C., *Children: Development & Relationships*, Macmillan Publishing Co., Inc., New York, 1977.
- Smithson, Michael, 'Possibility Theory, Fuzzy Logic, and Psychological Explanation' in Zétényi, Tamás (Ed.), *Fuzzy Sets in Psychology*, North-Holland, Amsterdam, 1988.
- Smolensky, Paul, *On the proper treatment of connectionism*, *Behavioral and Brain Sciences*, Vol. 11, Cambridge University Press, USA, 1988.
- Sneath, Peter H. A. and Sokal, Robert R., *Numerical Taxonomy, The Principles and Practice of Numerical Classification*, W. H. Freeman and Company, San Francisco, 1973.

- Sneath, Peter H. A., 'Philosophy and method in biological classification', in Felsenstein, J. (Ed)., *Numerical Taxonomy*, Springer-Verlag, Berlin, 1983.
- Somjen, George, *Neurophysiology - the essentials*, Williams & Wilkins, London, 1983.
- Sontag, Eduardo D., *Feedback Stabilization using Two-Hidden-Layer Nets*, Report SYNCON-90-11, Rutgers University, New Brunswick, New Jersey, October 1990.
- Sontag, Eduardo D., *Feedforward Nets for Interpolation and Classification*, SYCON - Centre for Systems and Control, Department of Mathematics, Rutgers University, New Brunswick, 1990.
- Steel, Robert. G. D. and Torrie, James H., *Principles and Procedures of Statistics, A Biometrical Approach*, Second Edition, McGraw-Hill Book Company, Singapore, 1987.
- Sternberg, Robert J. & Davidson, Janet E., (Eds.), *Conceptions of Giftedness*, Cambridge University Press, Cambridge, 1986.
- Sternberg, Robert J., 'General Intellectual Ability', in Sternberg, Robert J. & Davidson, Janet E., (Eds.), *Conceptions of Giftedness*.
- Stevens, Charles F., *The Neuron*, in *Progress in Neuroscience*, W.H. Freeman and Company, New York, 1986.
- Stirling, David, & Buntine, Wray, *Process Routings in a Steel Mill, a challenging induction problem*, New South Wales Institute of Technology, Broadway, N.S.W., 1987.
- Strickberger, Monroe W., *Genetics*, The Macmillan Company, New York, 1968.
- Sun, R., *Integrating Rules and Connectionism for Robust Reasoning*, Technical Report TR-CS-90-154, Brandeis University, Waltham, U.S.A., 1991.
- Sun, R., *Connectionist Models of Rule-Based Reasoning*, to appear in the Proceedings of the 13th Annual Conference of the Cognitive Science Society, 1991

- Szillard, Von L., 'Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen', in *Zeitschrift für Physik*, Volume 53, Verlag von Julius Springer, Berlin, 1929.
- Tesauro, Gerald, "Neural Models of Classical Conditioning: A Theoretical Viewpoint", in: Hanson, Stephen Jose and Olson, Carl R., *Connectionist Modelling and Brain Function: The Developing Interface*, The MIT Press, Cambridge, Massachusetts, 1990.
- Tesauro, Gerald, 'Temporal Difference Learning in Backgammon Strategy', in Sleeman, Derek and Edwards, Peter, *Machine Learning Proceedings of the Ninth International Workshop*, Morgan Kaufmann Incorporated, 1992.
- Thompson, Richard F., *The Brain*, W. H. Freeman and Company, New York, 1985.
- Thompson, Richard F., *Progress in Neuroscience*, W. H. Freeman and Company, New York, 1986.
- Thompson, Thomas M., *From Error-Correcting Codes Through Sphere Packings to Simple Groups*, The Mathematical Association of America, 1983.
- Trusted, Jennifer, *Free Will and Responsibility*, Oxford University Press, 1984.
- Tulving, E., "Cue-dependent forgetting", *American Scientist* 62, 1974, pps. 74-82 (not seen, referred to in Mayes, J. Terry, Draper, Stephen W., McGregor, Alison M. and Oatlet, Keith, "Information Flow in a User Interface: The Effect of Experience and Context on the Recall of MacWrite Screens", in Preece, Jenny and Keller, Laurie (Eds.), *Human-Computer Interaction*, Prentice Hall, Hertfordshire, England, 1989).
- Van de Velde, Walter, 'Incremental Induction of Topologically Minimal Trees', in Porter, Bruce and Mooney, Raymond (Eds.), *Machine Learning: Proceedings of the Seventh International Conference*, Morgan Kaufmann Publishers Inc., San Mateo, 1990.
- Vaux, Janet, "Replicating the expert", *New Scientist*, 3rd March 1990, pps 39-42.

- Vinacke, W. Edgar, *Foundations of Psychology*, American Book Company, New York, 1986.
- Vines, Gail, "The Emotional Chicken", *New Scientist*, 22 January, 1994, New Science Publications, London, 1994.
- Wakerly, John, *Error Detecting Codes, Self-Checking Circuits and Applications*, North-Holland, New York, 1978.
- Waldrop, M. Mitchell, *Complexity - the emerging science at the edge of order and chaos*, Penguin books, London, 1994,
- Watson, Robert I., & Lindgren, Henry Clay, *Psychology of the Child*, John Wiley & Sons, Inc., New York, 1959.
- Wechsler, David, *The Measurement and Appraisal of Adult Intelligence*, The Williams & Wilkins Company, Baltimore, U.S.A.1958.
- Werner, Gregory M. and Dyer, Michael G., *Evolution of Communication in Artificial Organisms*, Technical Report UCLA-AI-90-06, University of California, Los Angeles, USA, November 1990.
- Widrow, B. "Generalization and Information Storage in Networks of Adaline Neurons", in : Yovits, M.C., Jacobi, G.T., & Goldstein, D. (Eds.), *Self-Organizing Systems 1962*, Spartan Books, Washington DC, 1962, (not seen, quoted in Zeidenberg, Matthew, *Neural Networks in Artificial Intelligence*, Ellis Horwood, New York, 1990).
- Wierzbicki, A., 'Interactive decision analysis and interpretative computer intelligence', in *Interactive Decision Analysis*, Springer-Verlag, Berlin, 1984.
- Melsa, James L. and Cohn, David L., *Decision and Estimation Theory*, McGraw-Hill, New York, 1978.
- Williams, Graham J., *Some Experiments in Decision Tree Induction*, The Australian Computer Journal, Vol. 16, No. 2, May 1987.
- Williams, W. T., Dale, M. B. and Macnaughton-Smith, P., 'An Objective Method of Weighting in Similarity Analysis', *Nature*, January 25, 1964, p. 426.

Williams, W. T. and Lambert, J. M., 'Multivariate Methods in Plant Ecology', *The Journal of Ecology*, Volume 48, Blackwell Scientific Publications, Oxford, 1960, pps. 689-710.

Williams, W. T., Lambert, J. M. and Lance, G. N., 'Multivariate Methods in Plant Ecology', *The Journal of Ecology*, Volume 54, 1966, pps. 427-445.

Winograd, Terry, 'Computer Programs for Inductive Reasoning', in Cohen, Jonathon and Hesse, Mary (Ed.), *Applications of Inductive Logic*.

Winston, Patrick Henry, *Artificial Intelligence*, (edn. 1), Addison-Wesley Publishing Company, Reading, Mass., 1979.

Winston, Patrick Henry, *Artificial Intelligence*, (edn. 2), Addison-Wesley Publishing Company, Reading, Mass., 1984.

Wirth, Jarryl and Catlett, Jason, 'Experiments on the Costs and Benefits of Windowing in ID3', in Laird, John (Ed.), *Proceedings of the Fifth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Mateo, U.S.A., 1988.

Wong, M.A. and Lane, T., "A  $k$ th Nearest Neighbour Clustering Procedure", *Journal of the Royal Statistical Society*, Series B, 45, 1983, pps. 362-368, (not seen), referenced in SAS Institute Inc, *SAS/STAT User's Guide*, Release 6.03, SAS Institute Inc., Cary, NC, 1988.

Wong, M.A. and Schaak, C., "Using the  $k$ th Nearest Neighbour Clustering Procedure to Determine the Number of Subpopulations", *American Statistical Association 1982 Proceeding of the Statistical Computing Section*, 1982, pps. 40-48, (not seen), referenced in SAS Institute Inc, *SAS/STAT User's Guide*, Release 6.03, SAS Institute Inc., Cary, NC, 1988.

Wright, Frank Lloyd, quoted in Minsky, Marvin, *The Society of Mind*.

Yazdani, Masoud and Narayanan, Ajit (Eds.), *Artificial Intelligence: human effects*, Ellis Horwood Limited, Chichester, 1984.

Zadeh, Lofti A., 'Fuzzy Sets versus Probability', *Proceedings of the I.E.E.E.*, Volume 68, No. 3, March 1980.



Zeidenberg, Matthew, *Neural Networks in Artificial Intelligence*, Ellis Horwood, New York, 1990,

Zwahlen, Helmut. T., Hartmann, Andrea L., and Rangarajulu, Sudhakar L., 'Effects of rest breaks in continuous VDT work on visual and musculoskeletal comfort/discomfort and on performance,' in Salvendy, Gavriel, (Ed.), *Human-Computer Interaction*, Elsevier Science Publishers, Amsterdam, 1984.

# Appendix A

## Clustering Methodology

This appendix details results obtained by using clustering methodology to aid species or taxa identification. Section A.1 presents some of the basic concepts, and postulates the types of results which could occur if the combination of methodology and data was ideal. Section A.2 presents the results obtained using the *Acaena* and *Danthonia* data.<sup>1</sup> Section A.3 comments on the utility of the clustering methodology in this case.

### A.1 Discussion on Clustering.

Clustering methodologies have long been used in biology in analyses of species diversity and forming hierarchical and non-hierarchical classifications. One of the earliest known to the author of this thesis is László Orlóci's 1969 paper in *Nature*, where, after referring to five previous applications of Shannon's entropy-based information theory in this area, he details an alternative clustering approach implemented in three PDP-10 Basic programs INF1, INF2 and INF3.<sup>2</sup> The clustering approach was of interest because, as the SAS user's guide comments:

The purpose of cluster analysis is to place objects into groups or clusters suggested by the data, not defined a priori, such that the objects in a given cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar.<sup>3</sup>

In the past this type of grouping has been the basis of some methodologies used to separate botanical specimens into genera, the relationship often being expressed by the use of dendrograms or keys.<sup>4</sup> For this reason it was decided to include

---

<sup>1</sup>The SAS statistical package running on a Sun 4 was used to provide the clustering algorithms.

<sup>2</sup>Orlóci, László, 'Information Analysis of Structure in Biological Collections', *Nature*, Volume 223, August 2, 1969, pps. 483-484.

<sup>3</sup>See SAS Institute Inc, *SAS/STAT User's Guide, Release 6.03*, SAS Institute Inc., Cary, NC, 1988, p. 47.

<sup>4</sup>For example, see Figures 6.7 and 6.8 and the surrounding discussion on pps. 102-104 in Ferguson, Andrew, *Biochemical Systematics and Evolution*, John Wiley

clustering methodologies in the survey of alternative methodologies to be examined in this thesis. The data was stripped of the identifications already assigned by the experts and submitted to the methodology.<sup>1</sup> The resultant groupings produced by the clustering methodology were then compared with the expert's classifications.

Ideally for the purposes of the identification of botanical species or taxa, each cluster would be composed of only one species or taxa, and the number of clusters would equal the observed number of species or taxa.

#### A.1.1 Finding the Number of Clusters.

Ideally a clustering methodology would be able to indicate the "natural" number of clusters into which the data would fall. However in practice:

There are no satisfactory methods for determining the number of population clusters for any type of cluster analysis (Everitt 1979, 1980).<sup>2</sup>

There have been many attempts to solve this problem.

Perhaps the best approach to the number-of-clusters problem that has yet appeared is provided by Wong and Schaak (1982). The  $k$ th-nearest-neighbour clustering method developed by Wong and Lane (1983) is applied with varying values of  $k$ .<sup>3</sup>

The SAS package implements both this method and several others which have been proposed. It was used in this investigation.

---

and Sons, New York, 1980, (note that the difference between a phenogram and dendrogram is defined on p. 8 of this reference).

<sup>1</sup>For more detail of the programs used to translate the data, see section 4.7 b) & f) of this thesis. Note that in this case training and test data sets were not required, the whole data sets were translated (with the species information in a form which was not available to the clustering procedures).

<sup>2</sup>See SAS Institute Inc., p.80. For completeness, the two Everitt references are included in the bibliography.

<sup>3</sup>See SAS Institute Inc., p.81. For completeness, the Wong and Schaak, and Wong and Lane references are included in the bibliography.

### A.1.2 Poorly and well separated clusters.

While it would be hoped that each cluster would contain only one species or taxa, in practice clusters are sometimes composed of specimens predominantly from one species or taxa, while also containing isolated and perhaps anomalous or misclassified specimens from other species or taxa. It can also occur that the types of measurements taken do not allow the separation of species by clustering methodology, several species being clustered together. In both these cases the species are said to be "poorly separated" by the data observed. They are thus difficult to separate and identify, given the measurements obtained.

Since "poorly separated" is a somewhat amorphous and ill-defined term, it is perhaps best understood by use of an example. The example chosen employed Fisher's Iris data.<sup>1</sup>

The Iris data was clustered using Ward's method.<sup>2</sup> To help obtain a clear display with the clusters as widely separated as the method employed allowed, the results obtained by clustering were further subjected to a discriminant analysis using the clusters obtained by Ward's method as the classes for the purpose of the discriminant analysis. The results were then plotted using the first and second canonical components on the axes, see Figure 31.

The cluster indicated by the symbol 1 of Figure 31 can be said to be well-separated from the clusters formed by the symbols 2 and 3.<sup>3</sup> The clusters formed by the symbols 2 and 3 are said to be poorly separated from each other.

---

<sup>1</sup>This is a commonly used small data set which is classified as being poorly separated. It was originally presented in: Fisher, R.A. 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics*, 7, pps. 179-188, 1936 (not seen). This data is also presented in SAS Institute Inc., p. 332.

<sup>2</sup>In Ward's minimum-variance method, the distance between two clusters is the ANOVA sum of squares between the two clusters added up over all the variables. The method is strongly biased towards producing clusters with roughly the same number of observations, and is very sensitive to outliers. For more details, see SAS Institute Inc., p. 15.

<sup>3</sup>For another example of well-separated, compact clusters, see SAS Institute Inc., p.51.

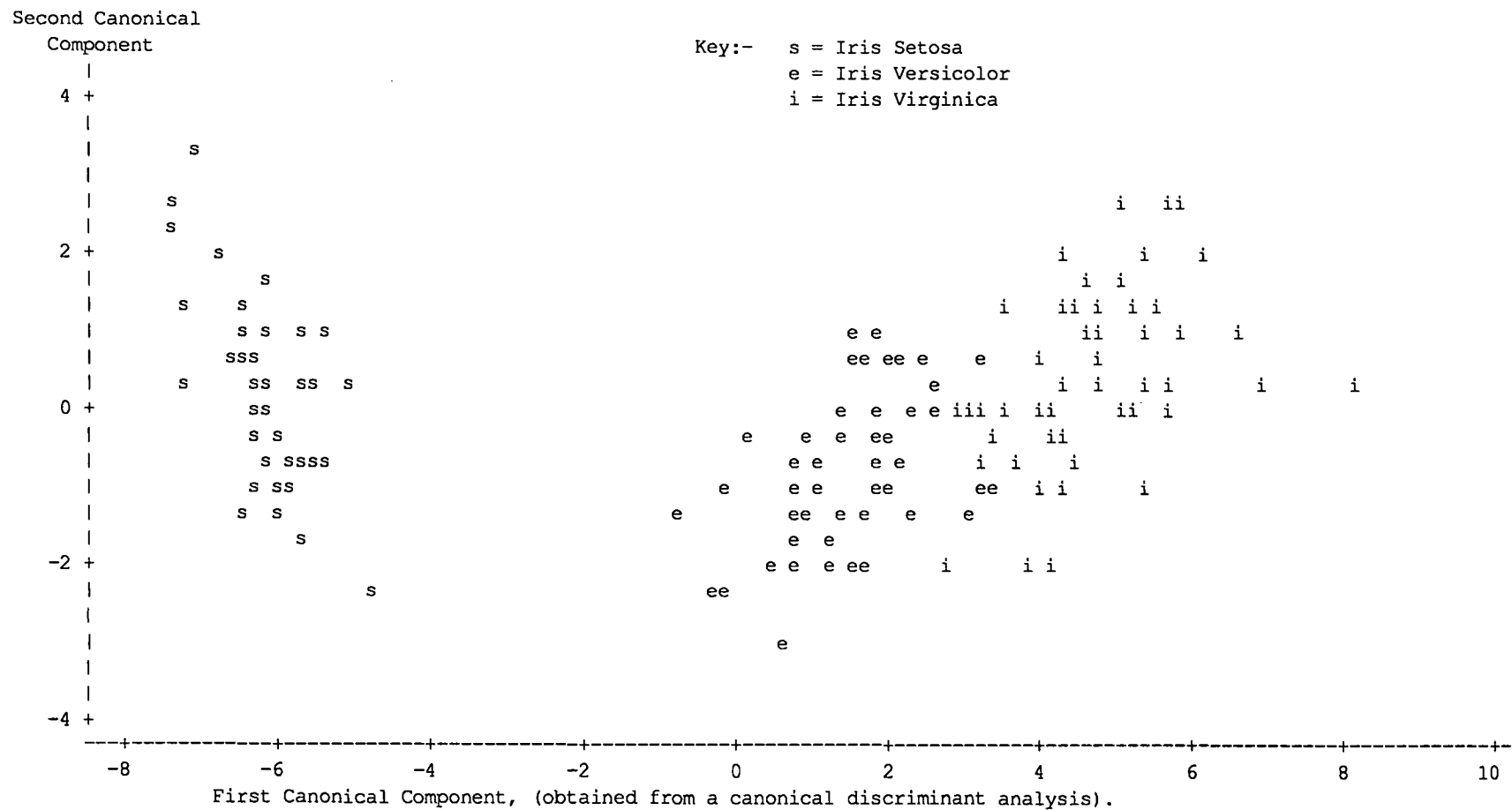


Figure 30 - Species-Specific Plot of Fisher's Iris Data, clustered by Ward's Method.

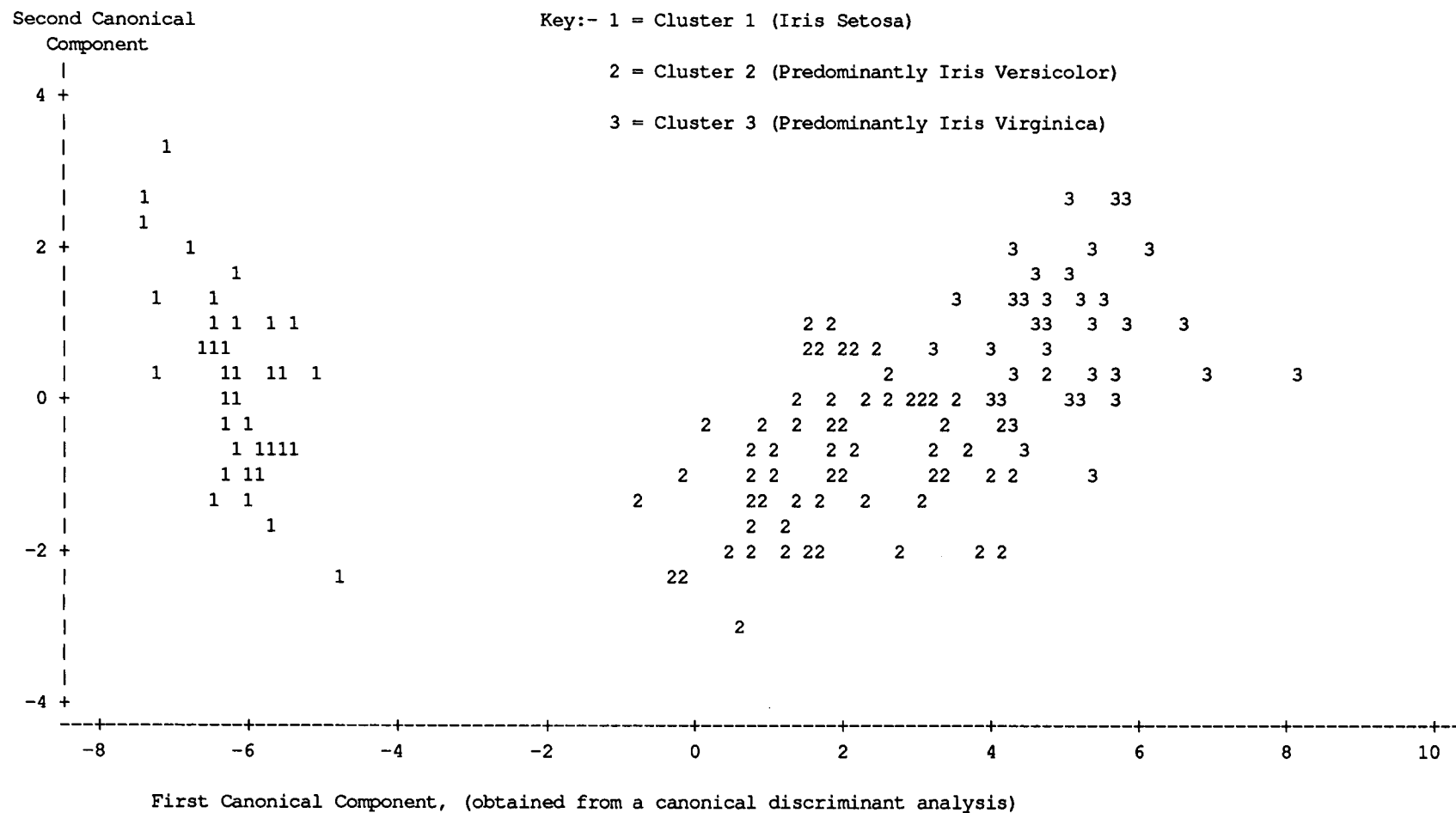


Figure 31 - Cluster-Specific Plot of Fisher's Iris Data, clustered by Ward's Method.

If Figures 30 and 31 are compared, it will also be noted that whilst cluster 1 is composed of one species (*Iris Setosa*), clusters 2 and 3 each contain more than one species. Cluster 2 predominantly represents specimens of *Iris Versicolor*, but also contains some specimens of *Iris Virginica*. Cluster 3 predominantly represents specimens of *Iris Virginica*, but also contains some specimens of *Iris Versicolor*, further confirming their classification as "poorly separated".<sup>1</sup>

### A.1.3 SAS clustering limitations involving incomplete data.

A further complication of the clustering methodology employed is the SAS clustering procedures requirement that all specimens have complete data, i.e. data must be supplied for each characteristic of each specimen of each species or taxa. Specimens which contain incomplete data will be rejected by the SAS clustering procedures.<sup>2</sup>

This requirement can be a significant limitation in the case of botanical specimens, where measurements are typically much more than usually prone to be incomplete because of the varying seasonality of many of the characteristics measured.

It was noted that Fisher's *Iris* data contained complete data.

## A.2 Results obtained using Clustering Methodology

Clustering methodology was applied to both the *Acaena* and *Danthonia* data.

Section A.2.1 lists the preliminary results obtained with the *Acaena* and *Danthonia* data using Ward's method of clustering plus discriminant analysis. This was used to get a preliminary look at the form of the data. Section A.2.2 details an attempt to see if there was a "natural" number of clusters in the *Acaena* and *Danthonia* data. Section A.2.3 gives the full results obtained for the *Acaena* data. Section A.2.4 gives the full results obtained for the *Danthonia* data.

---

<sup>1</sup>For other examples of compact clusters that are classified as poorly-separated, see SAS Institute Inc., p.53.

<sup>2</sup>See SAS Institute Inc., p.299.

The SAS requirement of complete data effected the two sets of data differently.

The *Danthonia* data include complete measurements of each characteristic.

Only 22% of the *Acaena* specimens had complete measurements, the rest having one or more items of data missing.<sup>1</sup> The requirement for complete data eliminated *Acaena echinata* var. *robusta*, as there were no complete sets of observations for this taxa.<sup>2</sup> It also reduced *Acaena agnipila* var. *protenta* to one specimen, and *Acaena echinata* var. *echinata* to 2 specimens. This would appear to make the *Acaena* data a very difficult set of data to examine by clustering techniques. However since the limitations of the *Acaena* data are not uncommon in botanical data, it was decided to proceed.

Ideally, the *Acaena* data would respond to this methodology with 10 clusters, each containing just one taxa. Similarly the hoped result of clustering the *Danthonia* data was separation into 19 clusters, each of which contained just one taxa.

#### A.2.1 Preliminary results using Ward's method plus discriminant analysis with the *Acaena* and *Danthonia* data.

Both the *Acaena* and *Danthonia* data were subject to clustering using Ward's method. A discriminant analysis was performed using the resultant clusters as the classes for the purpose of the discriminant analysis, and the results then plotted using the first and second canonical components on the axes; see Figures 32 and 33 (the *Acaena* data) and Figure 34 (the *Danthonia* data).

---

<sup>1</sup>Many characteristics (e.g. flowers) appear only in season, and the collector must be at the collecting site at exactly the right time of the year to obtain specimens exhibiting seasonal characteristics. Unfortunately, the varieties of the *Acaena* taxa extant in Australia occur in widely scattered parts of S.E. Australia, and collection of complete data has not been economically possible. (For a map showing the world-wide distribution of the genus *Acaena* see: Humphries, Christopher J. and Parenti, Lynne R., *Cladistic Biogeography*, Clarendon Press, Oxford, 1989, Figure 1.5, p. 6).

<sup>2</sup>Two characteristics for the taxa *Acaena echinata* var. *robusta* contain no data at all.



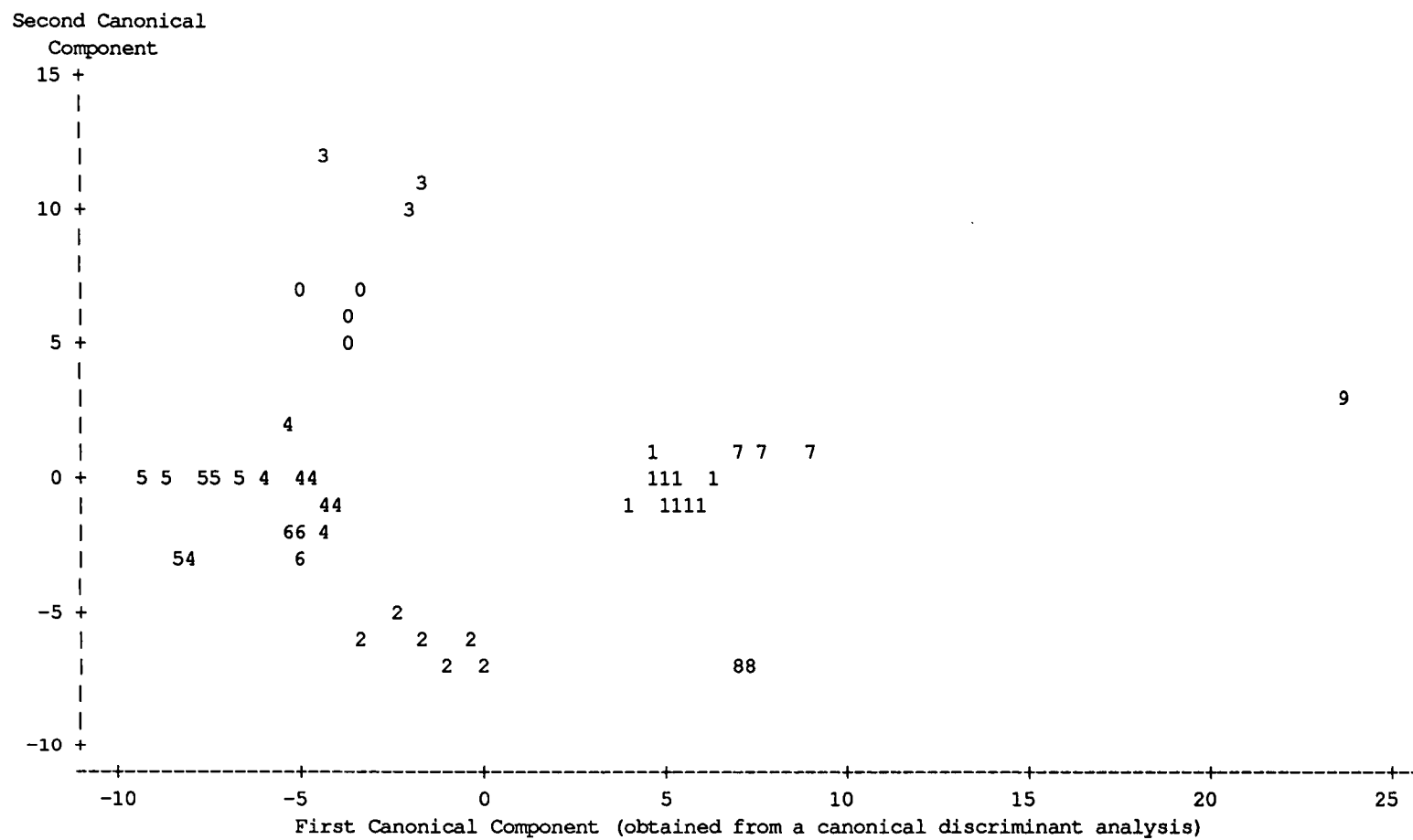
In the case of the *Acaena* data represented in Figure 32, the ten clusters are represented as the characters 0 to 9. Some of the clusters are well-separated, some poorly separated.<sup>1</sup> However, when the plot is compared with Figure 33, it can be seen that there is not always a one to one correspondence between species and cluster.

A complicating factor in this visual comparison is that not all of the specimens are represented on the plots. In some cases specimens of several species may occur very close to one another on the plot, and only one of the species may be represented by a letter on the plot.

This effect may be more easily seen in the case of the *Danthonia* data represented in Figure 34. Here the data is seen to form four broad groups in the plot, (which is plotted with the data grouped into only nine (instead of nineteen) clusters for clarity). In the left-most group on the plot, four different clusters are shown. A species-specific plot to the same scale shows six species, but a species-specific plot with doubled horizontal and vertical scales shows fourteen species. In fact, only about 1 in 10 of the specimens are represented directly on this plot, and thus the potential for missing species is large. To eliminate the effect of these concealed specimens (and hence species), subsequent results of the clustering runs will be presented as tables.

---

<sup>1</sup>The clustering was stopped at 10 clusters instead of 11 because the SAS requirement for complete data meant that one *Acaena* species was eliminated from consideration in the case of all the SAS runs.



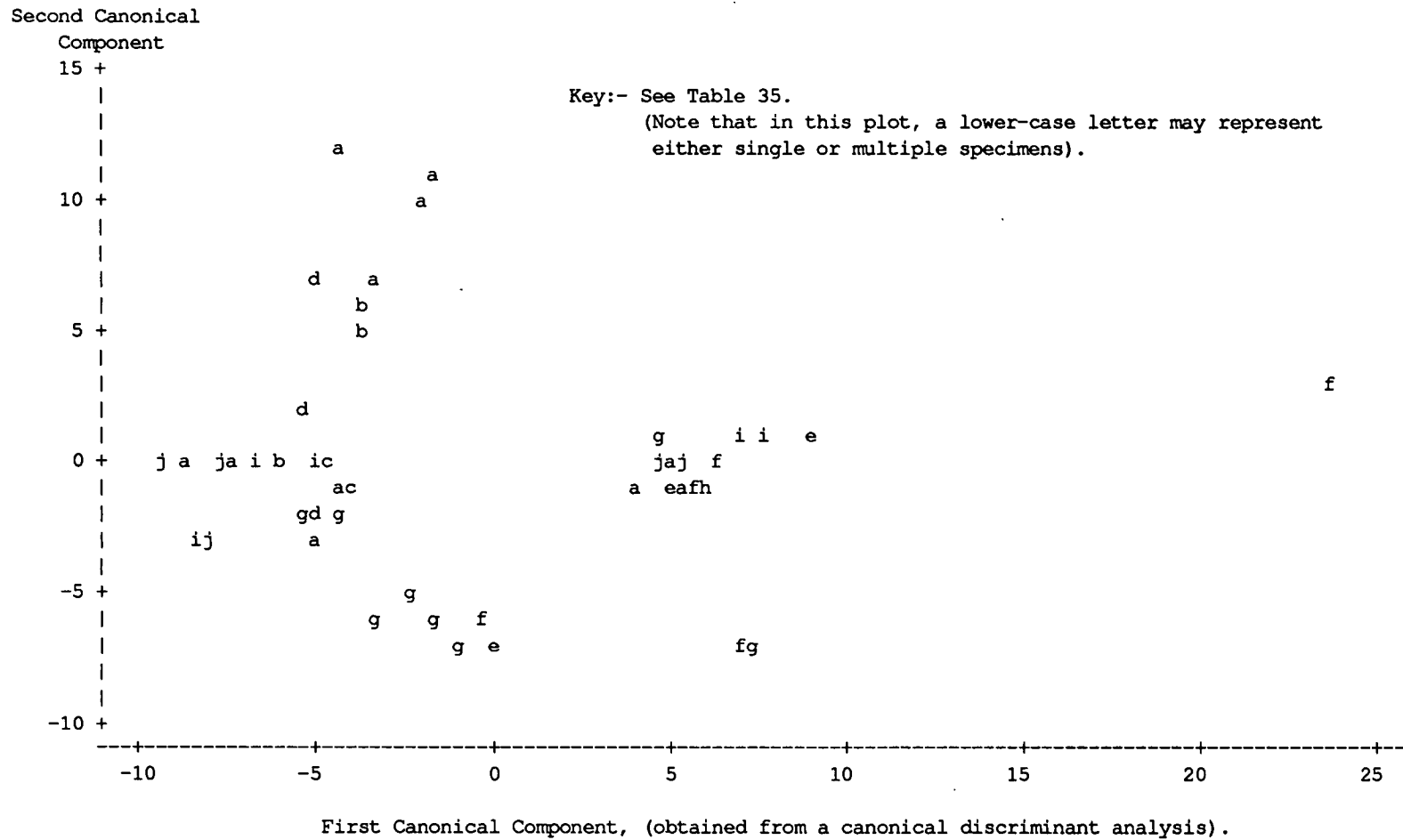


Figure 33 - Taxa-Specific Plot of the Acæna Data, clustered by Ward's Method.

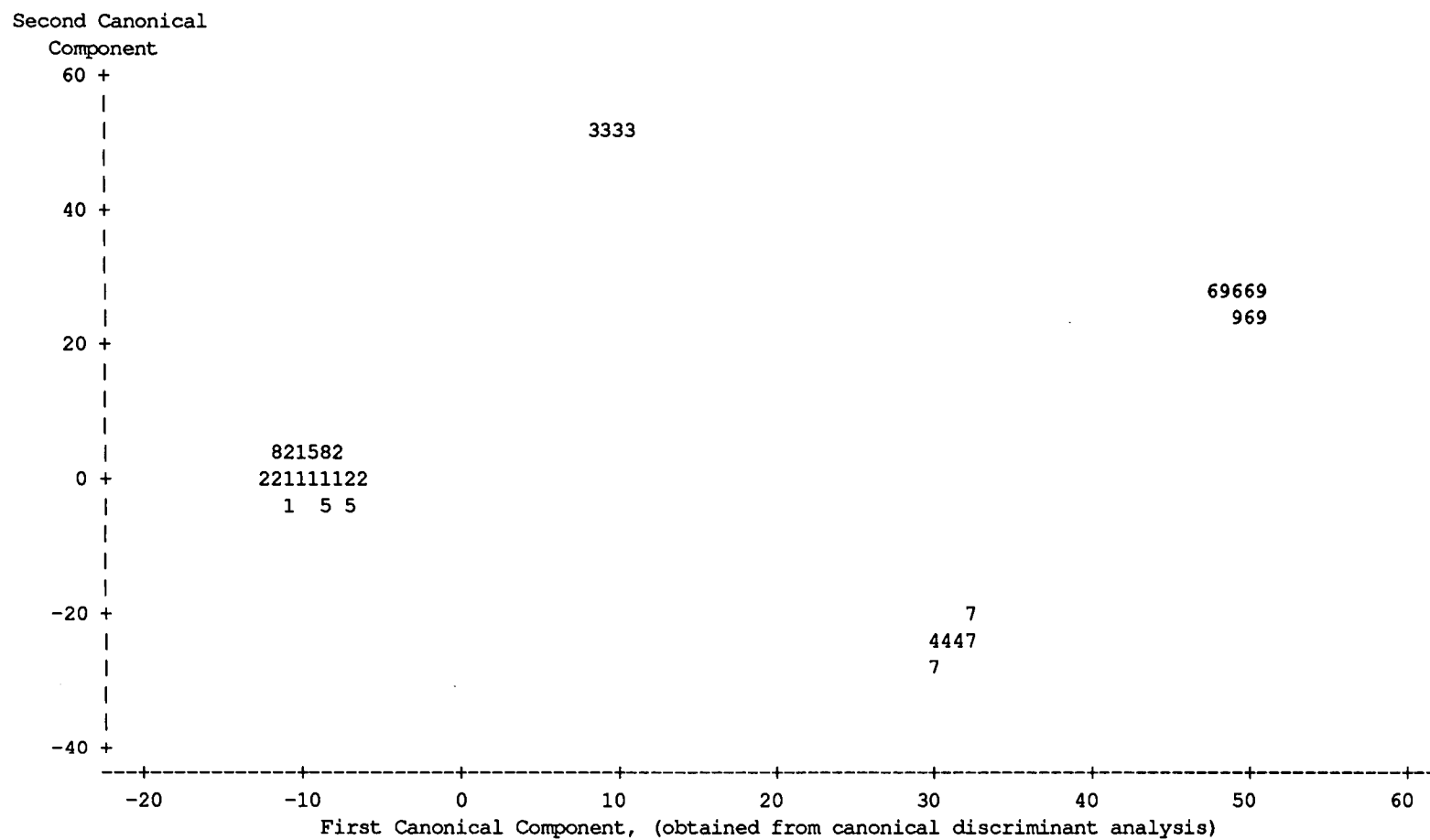


Figure 34 - Cluster-Specific Plot of Danthonia data, clustered by Ward's method.

### A.2.2 The number of clusters in the *Acaena* and *Danthonia* data.

Next an attempt was made to estimate the number of clusters into which the *Acaena* and *Danthonia* data “naturally” fitted.

The approach used initially was the *k*th-nearest-neighbour clustering method developed by Wong and Lane (1983), applied with varying values of *k*. ... This is implemented in the SAS CLUSTER procedure as METHOD = DENSITY with the K= option.<sup>1</sup>

Use of this methodology to attempt to determine the number of clusters did not produce clear results.<sup>2</sup>

Attempts involving over 30 other runs employing various combinations of different clustering methodologies and parameters followed, with little success.

The cause of the difficulty can be seen more clearly when the general method used to establish the clusters is examined. Roughly, the clustering approach involves treating each item of data as a singleton occupying a position in a (usually) multi-dimensional space. The “distance” between the singletons is then calculated according to some measure, the measure differing according to the method employed. The closest two singletons are then classified together to form a cluster of two specimens. This process is then repeated with the result being either a second cluster of two, or another specimen being joined to the originally-derived cluster to form a cluster of three specimens. The process then continues in a similar manner, with either singletons or other clusters being classified together.

It can be seen that this process, if continued to its logical conclusion, will eventually result in one huge cluster. Most methods attempt to avoid this largely useless eventuality by employing some sort of termination condition, which will stop the clustering process when the resultant clusters are judged to

---

<sup>1</sup>See SAS Institute Inc., p.81. For completeness, the Wong and Schaak, and Wong and Lane references are included in the bibliography.

<sup>2</sup>This method uses non-parametric probability density estimations, with the number of nearest neighbours being specified by the variable K. For more details, see SAS Institute Inc., pps. 293-294.

be distinct enough to make any further "joining" of the clusters or specimens counter-productive. This final number of clusters could be regarded as a measure of the "natural" number of clusters of specimens in the data, (according to the method of clustering and the termination conditions employed). However if the experimenter judged it appropriate, the number of clusters reached at some point before termination could also be used as the "natural" number of clusters.

As mentioned above, the results obtained with the *Acaena* and *Danthonia* data were not considered definitive. The results obtained with the *Danthonia* data could be used to broadly represent the results. Depending on the method of analysis used, results were obtained which suggested 1, 4, 7, 17, 32 or 100 "natural" clusters for the 19 species. Although two-stage clustering ( $k=4$ , number of modal clusters = 17) in the case of the *Danthonia* data was close, there was no consistent suggestion from the results obtained that there might be a natural clustering equal to the number of species for either data. The results of this investigation will not be presented in detail, as they are voluminous, vary widely, and do not add much information to the general picture represented in Figures 32 to 34.

As a result of this investigation it was noted that there was not a "natural" number of clusters which corresponded to the number of species or taxa in the *Acaena* and *Danthonia* data. It was decided that, if the clusters were to be allocated to specific species or taxa, some criteria other than pure clustering would have to be used.

Since there did not seem to be any "natural" clustering associating clusters with taxa directly, it was decided to attempt to "allocate" the clusters to taxa, to see what proportion of the data could be "identified" in this manner. This process would provide some estimate of the "separateness" of the taxa, and perhaps provide an additional insight into the difficulty of classifying the *Acaena* and *Danthonia* data. To do this, it was decided to stop the clustering at a point where the number of clusters was equal to the number of taxa being investigated. Runs which stopped above these limits were generally omitted from

further consideration.<sup>1</sup> Each cluster was then “allocated” to the taxa which had the highest proportion of its specimens amongst the specimens agglomerated into that cluster. In the event of a tie, the cluster was allocated to the species which had the highest number of specimens in the cluster.<sup>2</sup> In the event of this not resolving the tie, the cluster was allocated to the taxa which was not elsewhere associated with a cluster.<sup>3</sup> If a tie still remained, the cluster was allocated to a taxa by using a random number table.<sup>4</sup>

### A.2.3 Full clustering results using the *Acaena* data.

In the following sections of this appendix the results of the clustering runs are presented in the form of tables, with the clusters allocated to species or taxa using the method outlined in the previous section. Use of the full botanical name in the tables would take too much space, so in these tables each Taxa is represented by a single letter, as shown in Table 35.

On some occasions a species or taxa is represented in a cluster by only one (probably anomalous) specimen. To make the presence of these singleton outliers obvious, single specimens are represented by a lower case letter, whereas multiple specimens occurring in one cluster are indicated by an upper case letter.<sup>5</sup>

---

<sup>1</sup>E.g. *Danthonia* data with density (k=2) and two-stage density (k=2) runs, which suggested 100 clusters as a lower limit. However two runs of the *Acaena* data which only just exceeded the clusters = taxa limit were included, (see Tables 36 and 37, plus the comments that accompany these tables).

<sup>2</sup>On the basis that the higher number of specimens hopefully represented a more typical; sample of the taxa; (this is not, of course, certain, but in the absence of any other evidence seemed a reasonable hypothesis).

<sup>3</sup>This had the effect of enhancing the number of taxa represented, without diminishing the number of specimens “recognised” by this method.

<sup>4</sup>This had to be used only once in this investigation, to choose between two clusters each consisting of two specimens of differing taxa.

<sup>5</sup>A similar approach is used for the *Danthonia* data, see Table 41.

<b><i>Acaena</i> Taxa</b>	<b>If more than one specimen</b>	<b>If only one specimen</b>
<i>Acaena echinata</i> var. <i>subglabricalyx</i>	A	a
<i>Acaena echinata</i> var. <i>retrorsumpilosa</i>	B	b
<i>Acaena echinata</i> var. <i>echinata</i>	C	c
<i>Acaena echinata</i> var. <i>tylacantha</i>	D	d
<i>Acaena agnipila</i> var. <i>agnipila</i>	E	e
<i>Acaena agnipila</i> var. <i>tenuispica</i>	F	f
<i>Acaena agnipila</i> var. <i>aequispina</i>	G	g
<i>Acaena agnipila</i> var. <i>protenta</i>	H	h
<i>Acaena ovina</i> var. <i>ovina</i>	I	i
<i>Acaena ovina</i> var. <i>velutina</i>	J	j

Table 35 — Key to *Acaena* Taxa.

The allocated taxa is identified in the body of the tables by the use of a bold typeface.

The percentage of specimens occurring in clusters allocated to their taxa is noted at the bottom line of each of tables. For the purposes of these tables the figures are labelled “identified”, although some caution should obviously be used with the interpretation of this term.

The chance rate of identification of the *Acaena* taxa would be 9% if there was an equal number of specimens for each taxa represented in the data. This was not the case. The requirement for complete data eliminated one taxa (*Acaena echinata* var. *robusta*) from consideration, making the chance rate of identification 10% if there was an equal number of specimens for each of the remaining taxa represented in the data. This was also not the case. One taxa (*Acaena echinata* var. *subglabricalyx*) was represented by 23% of the complete specimens, and an observer with a knowledge of the data could obtain this percentage correct by guessing only this taxa.



The result of applying the Density Linkage cluster analysis to this data is shown in Table 36. It will be noted from this Table that using a value of  $K > 2$  in the Density Linkage clustering method led to a situation where almost all the data was collected into one huge cluster. This was expected to occur to at least some extent, as cutting the data being considered down to the 22% of specimens that had complete data meant that some taxa had two or less valid specimens, and  $K > 2$  would mean preferentially linking to a specimen of another taxa after the first link was made.<sup>1</sup> The  $K > 2$  runs were stopped at 10 clusters. The  $K = 2$  run formed 13 modal clusters, and this was accepted for inclusion in these results as an example of the type of variation which occurs when a higher clustering limit is accepted.<sup>2</sup>

---

<sup>1</sup> Assuming, of course, that the first link was to a specimen of the same taxa. This did not always occur.

<sup>2</sup> In the ultimate, 100% "identification" could be achieved by this method if (the number of clusters = the number of specimens). Hence it was suspected that using 13 clusters instead of 10 would increase the rate of "recognition".

CLUSTER	K=2	K=3	K=4
1	a,d,g	A,B,C,D,E,F,G,h,I,J	A,B,C,D,E,F,G,h,I,J
2	A,g,j	i	f
3	F,j	f	i
4	A,b,c,i	f	f
5	d,g,j	i	g
6	a,I,J	f	f
7	c,e,f,G	g	i
8	A	f	a
9	a,B,d	a	d
10	e,h	d	e
11	e,I	-	-
12	g,f	-	-
13	F	-	-
Identified	53%	30%	30%

Table 36 — Density Linkage Cluster Analysis for *Acaena* data.

Substituting Two-stage Density Linking clustering for the Density Linkage method was anticipated to reduce the tendency to group specimens into one huge cluster, because:

METHOD = TWOSTAGE is a modification of the density linkage that ensures that all points are assigned to modal clusters before the modal clusters are allowed to join.<sup>1</sup>

Table 37 represents the result of a run of this method. It may be noted that the K=2 run listed is almost identical to the corresponding run in Table 36. The tendency to cluster into one huge cluster noted in the Table 36 K>2 runs has, in fact, been

<sup>1</sup>SAS Institute, Inc., p. 296. By contrast, the original Density method allowed modal clusters to merge before all the outliers had been incorporated in the modal clusters.

reduced, however the percentage of specimens associated with specific clusters has not improved.

CLUSTER	K=2	K=3	K=4
1	a,d,g	A,c,E,F,G,h,i,J	A,b,C,d,E,F,G,h,i,J
2	A,g,j	A,B,d	A,B,D
3	F,j	a,b,c,d,g,i,j	A,I,J
4	A,b,c,i	A,I,J	f
5	d,g,j	i	g
6	a,I,J	f	f
7	c,e,f,G	g	i
8	A	f	f
9	a,B,d	f	i
10	e,h	d	a
11	e,I	-	-
12	g,f	-	-
13	F	-	-
Identified	53%	30%	30%

Table 37 — Two-stage Density Linkage Clustering for *Acaena* data.

Since Density Linkage is less effective at recovering compact clusters from small samples than are methods that always recover compact clusters<sup>1</sup> and since the *Acaena* data would qualify as “small samples” with only 22% of it’s data being available to the clustering procedures, other methods supplied by SAS were tried.

<sup>1</sup>SAS Institute Inc., p.294.

Jain *et al.* comment 'Several comparative studies ... conclude that Ward's method ... outperforms other hierarchical clustering methods'.<sup>1</sup> The results of applying it are shown in Table 38. Since Milligan is quoted as noting that Ward's method is particularly sensitive to outliers,<sup>2</sup> a run was also made using Ward's Method with 10% of the "worst" outliers removed. This resulted in only a small improvement, see Table 38.

Although the Single Linkage clustering method has a poor reputation,<sup>3</sup> it is theoretically known to be good at handling some types of irregularly shaped clusters, and so was also tried, see Table 38.<sup>4</sup> The Density (K>2) and Single Linkage methods tended to produce one huge cluster containing roughly  $\frac{3}{4}$  of the specimens, with the rest in 9 tiny other clusters.<sup>5</sup> The single linkage method is also known to be susceptible to this type of chaining. A run with a 10% trim of outliers produced a minimally better result, but not sufficiently improved to warrant presenting it here. The run using Ward's method with 10% outliers removed was slightly the best of the three runs in Table 38.

---

<sup>1</sup>Jain, Anil K. and Dubes, Richard C., *Algorithms for Clustering Data*, Prentice-Hall, New Jersey, 1988, p. 81.

<sup>2</sup>SAS Institute Inc., p.297. For completeness, the reference to Milligan's article is included in the bibliography to this thesis.

<sup>3</sup>"The method with the poorest overall performance has almost invariably been single linkage"; see SAS Institute Inc., p. 50.

<sup>4</sup>In the single linkage clustering method, the distance between two clusters is taken as the minimum distance between an observation in one cluster and an observation in another cluster. Because it does not impose an assumed shape on the cluster (also a benefit of the density methods), it theoretically should be useful in cases of elongated or irregularly shaped clusters. For further information about this method, see SAS Institute Inc., pps. 295-296.

<sup>5</sup>When the data with at least one characteristic missing were eliminated, some of the taxa were left with very few specimens representing them. This would have made the clustering process difficult for the higher values of K in the Density Linkage methods.

<b>CLUSTER</b>	<b>Single Linkage</b>	<b>Wards</b>	<b>Wards - 10% outliers</b>
1	A,b,C,d,E,F, G,h,I,J	a,B,d	a,B,d
2	f	g,e,h,F,A,J	e,h,F,j
3	g	G,e,f	G,e,f
4	e,i	A	A
5	f	g,C,a,b,d,i,j	g,C,a,b,d,i,j
6	d	A,I,J	A,I,J
7	A,B,d	g,a,d	g,A,j
8	a	e,I	g,a
9	i	g,f	e,I
10	f	F	a
Identified	32%	43%	45%

Table 38 — Single Linkage & Ward's Cluster Analyses for *Acaena* data.

SAS offered seven other methods of clustering data. They are the Average<sup>1</sup>, Complete<sup>2</sup>, EML<sup>3</sup>, Flexible<sup>4</sup>, McQuitty<sup>5</sup>, Centroid<sup>6</sup>

<sup>1</sup>This method clusters on the group average. Distances were squared in these runs. This method tends to produce clusters with small variances. For more information, see SAS Institute Inc., pps. 286, 292.

<sup>2</sup>In this method "the distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster. Complete linkage is strongly biased towards producing clusters with roughly equal diameters, and can be severely distorted by moderate outliers (Milligan 1980)" quoted from SAS Institute Inc., p.293. For completeness, the reference to Milligan's work is included in the bibliography of this thesis.

<sup>3</sup>This method "is similar to Ward's method, but removes the bias towards equal-sized clusters. Practical experience has indicated that EML is somewhat biased towards unequal-sized clusters", quoted from SAS Institute Inc., p.295. For further information, see this reference. This method took significantly longer than the other clustering methods employed in these investigations.

<sup>4</sup>The Flexible-Beta method was developed by Lance and Williams in 1967. The value of Beta may be specified by the user. Most runs were made with the default value of -0.25. Milligan's suggested value for Beta of -0.5 for data with many outliers was tried. There was a minor improvement, but not sufficient to warrant inclusion in these summary results. For more information on this method, see SAS Institute Inc., pps. 287, 295.

<sup>5</sup>McQuitty's method employs arithmetic averages combined with weighted average linkages. For more information, see SAS Institute Inc., pps. 286, 295.

and Median<sup>1</sup> methods. For completeness, runs were made with each of these. The results are presented in Tables 39 and 40.

CLUSTER	Average	Complete	Centroid	EML
1	a,B,d	g,A,B,D	A,B,d	a,B,d
2	g,e,h,F,A,J	e,h,F,a,j	g,e,h,F,A,J	g,e,h,F,A,J
3	G,e,f,c,a, d,j	G,e,f	G,e,f,C,b, a,d,i,j	G,e,f
4	c,b,A,I,J	A	A,I,J	A
5	g,a,d	g,C,b,a,d,i, j	g,a,d	g,C,b,a,d,i, j
6	e,I	A,I,J	e,I	A,I,J
7	g,f	g,A,j	f	g,a,d
8	a	e,I	g	e,I
9	f	g,f	f	g,f
10	f	F	f	F
Identified	38%	43%	30%	43%

Table 39 — Average, Complete, Centroid and EML Cluster Analyses of the *Acaena* data.

<sup>6</sup>This and the Median method have some superficial similarities, the Centroid method employing an unweighted pair group method using centroids, whereas the Median method uses a weighted pair group method using centroids. Distance data was squared in this example. For more information, see SAS Institute, pps. 286, 292.

<sup>1</sup>This uses centroids in the clusters, combined with a weighted pair-group method. Distance data was squared in this example. For more information see SAS Institute Inc., pps. 286, 295.

CLUSTER	Flexible	McQuitty	Median
1	B,a	A,B,d	A,B,d
2	g,e,h,F,A,J	g,e,h,F,A,J	g,e,h,F,A,J
3	G,e,f,j	G,e,f,c,A, d,i,j	G,e,f,j
4	A,d	c,b,A,I,J	g,C,A,b,d, I,J
5	C,A,b,d,i	G	e,I
6	a,I,J	e,I	g,f
7	g,a,d	g,f	d
8	e,I	d	f
9	g,f	f	f
10	F	f	i
Identified	40%	30%	36%

Table 40 — Flexible, McQuitty and Median Cluster Analyses, *Acaena* data.

The results presented above suggest that the *Acaena* data is not separated into clearly delineated clusters. It could reasonably be said to be even more poorly separated than Fisher’s Iris data, and thus should pose a significant challenge to classification algorithms.<sup>1</sup>

If this clustering methodology was used for the purposes of classification of the *Acaena* taxa, the results above would suggest that a rate of identification notably greater than chance could be achieved. However the variation in the obtained results between the methods employed would suggest that a careful preliminary investigation of the shape of the clusters would prove very useful.

<sup>1</sup>However note that other algorithms may have an advantage if they have some way of using that portion of the *Acaena* data which is incomplete.

#### A.2.4 Full clustering results for the *Danthonia* data.

Whereas the small numbers of some taxa in the *Acaena* data was expected to make this data a difficult challenge for the clustering methodology employed, it was hoped that the more complete *Danthonia* data, (which has no missing characteristic measurements), might be more amenable to an agglomeration analysis.

The *Danthonia* data was then considered. The chance rate of identification of the *Danthonia* species would be  $5\frac{1}{4}\%$  if there was an equal number of specimens for each species represented in the data. This was not the case. One species (*Danthonia caespitosa*) was represented by 9.6% of the specimens, and an observer with a knowledge of the data could obtain this percentage correct by guessing only this species.

Table 41 is the key used when presenting the results of the cluster analyses of this data.



<b>Danthonia Species</b>	<b>If more than one specimen</b>	<b>If only one specimen</b>
<i>Danthonia caespitosa</i>	A	a
<i>Danthonia carphoides</i> var. <i>angustior</i>	B	b
<i>Danthonia diemenica</i>	C	c
<i>Danthonia dimidiata</i>	D	d
<i>Danthonia fortuneae-hibernae</i>	E	e
<i>Danthonia geniculata</i>	F	f
<i>Danthonia gracilis</i>	G	g
<i>Danthonia laevis</i>	H	h
<i>Danthonia nitens</i>	I	i
<i>Danthonia nivicola</i>	J	j
<i>Danthonia nudiflora</i>	K	k
<i>Danthonia pauciflora</i>	L	l
<i>Danthonia penicillata</i>	M	m
<i>Danthonia pilosa</i>	N	n
<i>Danthonia procera</i>	O	o
<i>Danthonia racemosa</i>	P	p
<i>Danthonia semiannularis</i>	Q	q
<i>Danthonia setacea</i>	R	r
<i>Danthonia tenuior</i>	S	s

Table 41 — Key to *Danthonia* Species.

An examination of Tables 42 to 45 will show that, as with the *Acaena* data, clusters consisting of one species are rare.<sup>1</sup> Generally clusters consist of specimens from several species, although often with one species predominating, as is shown by the overall percentage “identification”.

<sup>1</sup>If one excluded “clusters” consisting of one specimen.

CLUSTER	EML	Ward's minus 10% outliers	Ward's
1	d,I,J,k,L	d,I,J,k,L	d,I,J,k,L
2	a,C,D,E,K, M,N,P,R,s	C,D,e,K,L, M,N,P,q,r,S	A,C,D,E,H,I,K,L,M ,N,P,Q,r,S
3	C,D,E,H,I, K,L,n,p,Q,r,S	c,D,E,H,I, K,L,n,q	B,E,F
4	B,E,F	B,E,F	A,G,n,O,p,Q,R,S
5	A,C,d,H,K,M,N, P,s	A,C,H,K,m,p	C,d,H,K,M,N,P,s
6	A,G,n,O,p,Q,R,S	A,n,q,R,s	A,e,k,M,N,P,Q,R
7	A,e,k,M,N,P,Q, R	a,g,o,p,Q,R,s	c,E
8	a,B,F	C,D,H,K,M, N,P,R,s	a,C,D,K,M,N, P,R,S
9	a,B,q,R,S	A,c,e,K,M,N, P,Q,R,s	a,B,F
10	A,g,h,M,N,O,P, Q,r	c,E	a,B,q,R,S
11	a,d,h,M,N,P,r,S	a,B,F	A,g,h,M,N,O, P,Q,r
12	a,F,G	a,B,q,R,S	a,d,h,M,N,P,r,S
13	c,d,k,N,p,S	A,H,M,N,O,P,q,r	a,F,G
14	A,Q,R	c,k,N,p,S	A,Q,R
15	D,k	G	D,k
16	a,M,O	A,Q,r	a,M,O
17	G	D,k	G
18	A,h,p,s	d,h,M,N,P,r,s	A,h,p,s
19	a	F,g	a
Identified	35%	38%	36%

Table 42 — EML and Ward's Cluster Analyses, *Danthonia* data.

CLUSTER	Average	Complete	McQuitty	Flexible
1	A,C,D,E,H, I,J,K,L,M, N,P,Q,R,S	C,d,E,H,I,J, K,L,n,q	A,C,D,E,H,I, K,L,M,N,P,Q, R,S	d,I,J,k,L
2	B,E,F	A,C,D,E,H,K, L,M,N,P,Q,R,S	B,E,F	A,C,D,H,K, M,N,P,Q,R,S
3	A,G,n,O,p, Q,R,S	B,E,F	A,B,n,Q,R,S	C,D,e,H,I, K,L,n,p,Q,r,S
4	A,c,d,h,K,M ,N,P,Q,R,S	A,g,n,R,s	A,G,O,p,Q,R,s	c,E,l
5	A,B,F,g	A,G,O,p,Q,R,s	c,K,M,N,O,P, Q,r,S	B,E,F
6	a,B,Q,R,S	A,C,e,H,k, M,N,P,Q,R,S	A,B,F,g	A,G,n,R,S
7	A,e,g,h,M, N,O,P,Q,R	a,B,F	A,g,N,Q,R	A,G,O,p,Q,R
8	f,G	a,B,Q,R,S	f,G	A,B,F,g
9	A,n,Q,R	A,g,h,M,N, O,P,Q,r,s	D,k	a,B,Q,R,S
10	D,k	a,F,G	A,M,n,O,p	A,e,g,h,M, N,o,P,Q,R,s
11	A,d,h,p,s	A,r	Q,r	C,D,K,m,N,P,S
12	D	A,Q,r	D,N,p,r	a,d,h,M,N,P, r,S
13	m,o	a,h,M,n,O,p,Q	A,h,N,o,p	f,G
14	a	D,k,N	A,d,h,s	A,Q,R
15	q	f,G	a,g	a,M,O,q
16	a	A,d,N,P,r	g	A,M,n,O,p
17	a	A	a	A,d,h,N,p,s
18	g	o	a	A
19	o	g	o	g
Identified	28%	29%	30%	37%

Table 43 — Average, Complete, McQuitty and Flexible Cluster Analyses, *Danthonia* data.

CLUSTER	Density (K=4)	Density Two- stage (K=3)	Density Two- stage (K=4)	Density Two- stage (K=5)
1	A,C,D,E,g,H, I,J,K,L,M,N, O,P,Q,R,S	A,C,D,E,H,I, J,K,L,M,N,P, Q,R,S	d,I,J,K,l	d,I,J,K,l
2	A,B,G,n,O,p, Q,R,S	C	c,h,i,L,q	c,E,h,i,L,q
3	A,B,F,G	a,B,F	L	A,C,D,E,H,K, L,M,N,P,Q,R,S
4	B,E,f	A,B,n,o,q,R,s	A,C,D,h,K,M, N,o,P,Q,S	d,K,M,N,p
5	f	a,b,q,R,S	A,C,D,F,H,k,L M,N,P,q,R,S	c,d,I,K,M,N,q
6	f	m,n,p	E	A,b,G,n,o,q,R
7	a	A,g,h,N,O,Q,r	d,I,K,N,P,q	a,B,q,R,S
8	a	A,c,N,p,S	c,E	A,g,h,M,N,O, P,Q,R
9	q	B,E	C,H,N	A,B,F,G
10	a	A,M,n,O,P,Q, r	A,B,G,n,o,q, R,S	B,E,F
11	g	a,p,q,R	a,b,q,R,S	a,G,O,P,Q,R
12	d	M,N,p,s	A,B,F,G	a
13	o	a,F,G	A,g,M,N,P,Q,R	m
14	g	g,M	A,M,n,O,P,Q,R	g
15	m	G,o,p,Q,R,s	d,h,M,N,p,s	o
16	o	A,Q,r	B,E,F	o
17	a	e,F	G,o,p,Q,R	a
18	g	A,h,o,p,s	g	g
19	a	G	a	a
Identified	20%	25%	45%	35%

Table 44 — Density and Two-stage Density Analyses, *Danthonia* data.

CLUSTER	Centroid	Single Linkage	Median
1	c,d,h,I, <b>J</b> ,K,L	A,C,D,E,g,H,I, <b>J</b> ,K, <b>L</b> ,M,N,O, <b>P</b> , <b>Q</b> ,R,S	c,d,E,H,I, <b>J</b> ,K, L,n,q
2	A,C,D,E,H,I, <b>K</b> , L,M,N,P, <b>Q</b> ,R, S	<b>B</b> , <b>E</b> ,	<b>C</b> ,D,e,I,K,L,M, N,P,q,r,S
3	<b>B</b> ,E,F	A,B,G,n,O,p, <b>Q</b> , <b>R</b> ,S	A,C,d, <b>H</b> ,K,M, N,P, <b>Q</b> ,R,S
4	A, <b>G</b> ,n,O,p, <b>Q</b> , R,S	A,B, <b>F</b> ,G	<b>B</b> ,E,F
5	A,B, <b>F</b> ,g	<b>G</b>	A,B,g,n, <b>Q</b> , <b>R</b> ,S
6	a,B, <b>Q</b> , <b>R</b> ,S	e, <b>F</b>	A,G, <b>O</b> ,p, <b>Q</b> ,R,s
7	A,g,h,M,N, <b>O</b> , P, <b>Q</b> ,R,s	<b>a</b>	a,C,D, <b>E</b> ,H,K, M,N,P,S
8	f,G	<b>p</b>	A,B, <b>F</b> ,g
9	<b>d</b> ,k	<b>d</b>	A,g,H,M,N,O, P, <b>Q</b> ,R,S
10	<b>A</b> ,d,h,n,p,s	<b>a</b>	f, <b>G</b>
11	<b>D</b>	<b>d</b>	A, <b>Q</b> ,r
12	m, <b>o</b>	<b>q</b>	<b>D</b> ,N,p,r
13	<b>q</b>	<b>a</b>	m, <b>o</b>
14	<b>a</b>	<b>m</b>	<b>a</b>
15	<b>a</b>	<b>o</b>	<b>a</b>
16	<b>q</b>	<b>o</b>	<b>o</b>
17	<b>a</b>	<b>a</b>	<b>a</b>
18	<b>o</b>	<b>g</b>	<b>a</b>
19	<b>g</b>	<b>a</b>	<b>g</b>
Identified	25%	21%	33%

Table 45 — Centroid, Single Linkage and Median Cluster Analyses, *Danthonia* data.

The 'identification' rate is in every case well above the chance percentage. These methods could be used to significantly improve the rate of identification of species or taxa above that of guessing, but is less than ideal in its requirements for complete data.

### A.3 Summary

Clustering methodology produced rates of identification superior to that achievable on average by chance for both the *Acaena* and *Danthonia* data.

The rate of identification achieved in the case of the (complete) *Danthonia* data, although mostly lower in numerical terms than the (incomplete) *Acaena* data results,<sup>1</sup> is proportionally better than the rate noted for the *Acaena* data if one takes note of the expected chance identification.<sup>2</sup> This would appear to confirm the difficulty incomplete data caused to the clustering methodology used.

Between methods, the rate of identification varies widely, from 20% to 45%, and the clustering methodology which gave the equal highest rate with the *Acaena* data (density, K=2, 45%) gave the lowest rate on the *Danthonia* data (density, K=4, 20%). In both cases, the most appropriate clustering methodology would appear to be dependant on the shape of the clusters which naturally occur in the data.

In summary, clustering methodology would appear to be a useful, albeit limited methodology in the classification of botanical species and taxa.

---

<sup>1</sup>Although in one case the *Danthonia* "identification" rate actually exceeds the corresponding *Acaena* rate, (two-stage density clustering, K=4, see Tables 37 and 44), and is equal in another case, (McQuitty clustering, see Tables 40 and 43).

<sup>2</sup>See the discussion in section A.2.3 and section A.2.4.

# Appendix B: Neural Networks

This Appendix examines an alternative methodology which can be used for the classification of botanical specimens, neural networks. In this case data is split into two portions, the neural net trained on one set of data, and then the trained net is used to attempt to identify the second portion of the data.

Because the neural net techniques have only recently been revived and may be unfamiliar, the background to this important technique is covered in more detail than is the case of the other alternative approaches examined in Appendices A, C, D and E.

Neural net methodologies originated in attempts to imitate functions of the human brain. The methodologies assume (controversially) that the brain is a tabula rasa, written on by the experience gained from the learning data.<sup>1</sup> Section B.1 notes opinions expressed about such attempts. This section also notes that the brain is made up of many individual neurons richly connected into a network. Section B.2 presents an attempt to model a single neuron. Section B.3 comments on ways that have been proposed to join these model neurons into networks. Section B.4 covers some theory of the type of network chosen for use in this study. Section B.5 looks at implementation issues. Section B.6 presents the results obtained. Section B.7 discusses these results. Section B.8 presents a summary.

## B.1 Can the human brain be imitated?

The human brain is the most complex structure in the known universe.<sup>2</sup>

The brain has interested computer scientists, (amongst others), because it has permitted *Homo sapiens sapiens* to exhibit remarkable abilities in the fields of pattern recognition, reasoning, problem solving and language ability (to name a few). Also in some cases useful results appeared to have been obtained with little formal training being available. Neural net methods

---

<sup>1</sup>Remelhart et. al., p. 278. Also see Cromer p. 185.

<sup>2</sup>Thompson, Richard F., *The Brain*, W. H. Freeman and Company, New York, 1985, p. 1.

originated from attempts to mimic the approach thought to be used by the human brain, in an attempt to see if imitation could help reproduce some of these perceived abilities.

In the following discussion section B.1.1 comments on the traditional methods used by computer science to produce "intelligent" programs. Usually these approaches (e.g. expert systems) use rules, and section B.1.2 queries whether rules are necessary, or if intelligence is a function of the architecture used. Section B.1.3 notes the different approaches used to investigate intelligent behaviour, some using a "bottom-up" approach, using robots or artificial "insects", others using a "top-down" approach by investigating humans. Section B.1.4 notes reasons for the recent interest in imitating aspects of the brain's massively parallel computing paradigm. Section B.1.5 looks at attempts to understand how the brain works, including the psychologist's "black box" and the neuroanatomist's "divide-and-conquer" approaches. Section B.1.5.3 notes an attempt to combine these approaches.

### B.1.1 Attempts to imitate the brain.

There have been previous attempts to imitate these abilities. Computer methods using the traditional 'glorified adding machine'<sup>1</sup> Von Neumann architecture (used in all commonly available computers) have produced results which, whilst being remarkable in computer terms, are often unexceptional in biological terms.<sup>2</sup> This may be partially because current computers need a codex of algorithmically defined rules.<sup>3</sup> Only

---

<sup>1</sup>Hecht-Nielsen, Robert, *Neurocomputing: picking the human brain*, IEEE Spectrum 25(3), March 1988, p. 36.

<sup>2</sup>For example, Sejnowski and Churchland comment that 'The most powerful of today's computers approach speeds of 10 GFLOPS (1 billion operations per second)', but that 'A honeybee's brain, roughly and conservatively, performs at about 10 TFLOPS (10,000 GFLOPS)'. Also 'A honeybee's brain dissipates less than 10 microvolts. It is superior by about 7 orders of magnitude to the most efficient of today's manufactured computers.'; see: Sejnowski, T. and Churchland, P., 'Silicon Brains' in *Australian Personal Computer*, Vol. 13 No. 11, Computer Publications Pty. Ltd., November 1992, p. 136.

<sup>3</sup>They follow the type of picture first postulated for the brain by John Hughlings-Jackson (born 1835), who 'opted for a hierarchical principle that reflected ideas [strongly proposed by Herbert Spencer] about the evolution of species and the development of civilisation' (Ferry, p. 41). Spencer's ideas were later generalised by Charles Darwin. 'In the latter half of the 20th century, hierarchies of dominance have fallen from favour in the eyes of social psychologists and biologists alike. ... The new metaphor for the age is ... a set of logical rules for interaction between interdependent components' (Ferry p. 42). Computer architects have only just started to follow this path, e.g. see other references in



some humans trained in this methodology work this way, it seems some never do.<sup>1</sup>

### B.1.2 Are rules necessary?

It would be unusual to use a rule when e.g. recognising our father: *this human has shorter hair than that human, hence it must be my father, not my mother*. Frank Lloyd Wright said 'An expert is one who does not have to think. He knows.'<sup>2</sup> Much human knowledge is like this, absorbed from our surroundings without consciously forming rules.

The idea that intelligence might be a feature of something other than rules, e.g. an architecture, is noted by Massaro who contrasts 'a system whose architecture enables it to respond to inputs and one whose architecture provides it with the ability to use rules...'<sup>3</sup> Searle goes further, suggesting that the architecture may well have to be biological, as it is biology that matters.<sup>4</sup> Massaro notes that Boden rejects this view, stating that the brain 'as an organ of intelligence almost certainly has nothing to do *with what it is made of*. Rather, they concern *how it is organised and what it does*';<sup>5</sup> and 'Functions are where it's at: protoplasm has nothing, essentially, to do with it'<sup>6</sup> Boden's position is essentially that of a functionalist, i.e. 'that mental states can be characterised abstractly from whatever physically realises them (neural systems or silicon chips)'<sup>7</sup> Minsky is of similar mind, rejecting the ideas postulated by Searle and Roger

Penrose who says he believes the brain uses non-algorithmic and non-computational mechanisms when it comes to making conscious judgements ... I believe Roger Penrose uses low

---

this thesis to the computers produced by the Thinking Machines Corporation, and the dataflow computer; see Koppel, Tom, 'Profile: Supercomputer Solo', *Scientific American*, Volume 264, Number 3, March 1991. pps. 16-17.

<sup>1</sup>See previous discussion in chapter 1 of this thesis.

<sup>2</sup>Frank Lloyd Wright, quoted in Minsky, Marvin, *The Society of Mind*, Simon and Schuster, New York, 1986, p. 137.

<sup>3</sup>Massaro, Dominic W., Book Review, *American Journal of Psychology*, Vol. 104, No. 2, p. 282, Summer 1991.

<sup>4</sup>Searle, J.R., *Is the Brain's Mind a Computer Program?*, *Scientific American*, Vol. 262, No. 1, January 1990, pps. 26-31.

<sup>5</sup>Boden, Margaret A., *Artificial Intelligence in Psychology: Interdisciplinary Essays*, Bradford Books, MIT Press, Cambridge, U.S.A., 1989, p. 47 (italics in the original, not seen), quoted in Massaro.

<sup>6</sup>*Idem.*, p. 58.

<sup>7</sup>Massaro, p. 279.

quality algorithms when he makes judgements about consciousness - Marvin Minsky<sup>1</sup>

### B.1.3 Should imitations start bottom-up or top-down?

Brookes comments that realisation is growing that 'traditional Artificial Intelligence offers solutions to intelligence which bear almost no resemblance at all to how biological systems work'<sup>2</sup>, and is working towards developing a view of intelligence from the "bottom up", studying robots, (rather than emphasising the "top down" decompositional approach, studying or philosophising about humans). He comments that in this case 'Intelligence is determined by the dynamics of interaction with the world'<sup>3</sup> Neural net architectures learn in a similar way, by interaction with their 'world' of data, building up a 'knowledge base' in the form of weights dependant on both their architecture and the data.<sup>4</sup>

### B.1.4 Imitating the brain's parallel processing.

The difficulty in obtaining increased performance from super computers using the Von Neumann architecture has also been a causative factor in the recent resurgence of interest in neural nets. The brain uses huge numbers of massively parallel but (in computing terms) very slow computing elements.<sup>5</sup> A similar type

---

<sup>1</sup>Minsky, Marvin, quoted in Beynon, David, *Father of AI blasts the 'philosophers'*, Computerworld, September 6, 1991, p. 10.

<sup>2</sup>Brooks, Rodney A., *Intelligence Without Reason*, Proceedings of the Twelfth International Conference on Artificial Intelligence, Volume 2, August 1991, p. 569.

<sup>3</sup>*Idem.*, p. 584.

<sup>4</sup>This type of knowledge has been criticised because it is unavailable for independent examination in the form of rules, however some preliminary progress has been made towards the representation of this type of knowledge in the form of rules, e.g. see: Sestito, Sabrina and Dillon, Tharam, *Using neural networks for the extraction of high level knowledge representations for machine learning*, Technical Report No. 5/89, May, 1989, Department of Computer Science, LaTrobe University, Victoria, Australia, 3083. There have also been attempts to combine rule-based and connectionist reasoning, e.g. Sun, R., *Integrating Rules and Connectionism for Robust Reasoning*, Technical Report TR-CS-90-154, Brandeis University, Waltham, U.S.A., 1991; also Sun, R., *Connectionist Models of Rule-Based Reasoning*, to appear in the Proceedings of the 13th Annual Conference of the Cognitive Science Society, 1991, (a revised version of the previous technical report).

<sup>5</sup>The speed of conduction of pulses in the neurons is about 5 metres per second in fine neurons up to about 125 metres per second in large ones. The time for the pulse to be conducted along the length of a neuron is about 0.3 milliseconds. After a neuron fires there is an absolute refractory period of about 10 milliseconds during which the neuron cannot fire again. There is also a relative refractory period of increased threshold. See Fishbach, Gerald D., 'Mind and

of architecture using many faster, electronic CPUs would seem to offer promise, but how to connect the CPUs?<sup>1</sup> Blakemore comments that the human visual

cortex seems to be capable, by virtue of its developmental "plasticity", to regulate its own input and adjust the properties of its cells receptive fields. This process may play an essential role in the establishment of the highly efficient parallel computing array that constitutes the adult visual cortex.<sup>2</sup>

Generally this degree of self-adjusting plasticity has not been available at the computer hardware level, but even so a study of biologically optimised computing mechanisms has helped. Some computers developed using approximations of biological models have proven both cheaper and faster than current machines,<sup>3</sup>

---

Brain', *Scientific American*, Vol. 267 No. 3, September 1992, p. 26; also Block, H.D., *The Perceptron, A Model for Brain Functioning*, in *Review of Modern Physics*, 34(1), January 1962, p. 124. Regarding operating times of the brain as a whole, Harth comments that all 'higher' brain functions extend over periods longer than 300 ms; see Harth, Erich, *Order and Chaos in Neural systems: An Approach to the Dynamics of Higher Brain Functions*, IEEE Transactions on Systems, Man and Cybernetics, Vol SMC-13, No. 5, September/October, 1983, p. 783. Regarding memory capacity, Kohonen suggests a 'group' of 500 neurons has  $10^6$  inputs by which it can be encoded, and  $10^8$  patterns would suffice to store one sensory experience every 10 seconds of a person's waking-state life; see: Kohonen, Teuvo, *Associative Memory*, Springer-Verlag, Berlin, 1977, p. 146.

<sup>1</sup>Much of the study of biological computing mechanisms has been performed on the "lower" animals, e.g. Bach, Ivan N., (*Data Complexity*, in *Neuron-Digest* Vol-5-no-51, December, 1989) comments that Figure 2.13 of page 23 of the DARPA Neural Network Study provides a comparison of the storage and speed of a leech, worm, fly etc.. Cliff (p. 21) strongly supports this approach, quoting Brookes (p. 7) and Hoyle (p. 17) who suggest use of the arthropods in an evolutionist and antianthropocentric rejection of the phylogenetically top-down study of intelligence. However since this is not the usual approach used by other writers (e.g. Fukushima), we will here make the observation that the human brain uses about  $10^{11}$  neurons, each neuron being connected to about  $10^3$  other neurons, giving a storage capacity of about  $10^{14}$  interconnects, and a speed of about  $10^{16}$  interconnections per second (Bach). To design a VLSI implementation using parallel CPUs with this massive degree of connectivity represents a significant challenge.

<sup>2</sup>Blakemore, Colin, 'Computational Principles of the Visual Cortex', *The Psychologist*, Vol. 4, No. 2, February 1991, p. 73; see also Shatz, Carla J., 'The Developing Brain', *Scientific American*, Vol. 267 No. 3, September 1992, pps. 34-41.

<sup>3</sup>In very approximate orders of magnitude, in 1989 1 MIP. on a mainframe cost about \$US100,000, 1 MIP. on a PC cost about \$US7,000, 1 MIP. on a transputer (designed for parallel processing) cost about \$US300. A Cray X-MP operating at about 0.7 gigaflops using 4 processors cost about \$US10M; a Connection Machine operating at about 2.5 gigaflops using 65,536 processors arranged in a hypercube architecture (each CPU being slower than the CPU of many present-day PCs) cost about \$US3M. (The bargain still appears to be the human brain, which Minsky stated (Nov. 89) is equivalent in capacity to 200 Connection Machines (Model CM-2). The former is both easier and more enjoyable to manufacture than the previously-mentioned computers, costs about \$Aus700,000 (1989) to bring to a tertiary level of performance, but suffers the major disadvantage that there is

and have produced results which seem to be superior to those produced by programs written in algorithmic languages, (particularly in the application area of pattern recognition, where neural net computers can produce results without having to be fed rules by programmers or knowledge engineers).<sup>1</sup> The neural net methodology is preferable in some cases even when a neural net is simulated in software on a Von Neumann architecture computer.

### B.1.5 How does the brain work?

Finding how the brain works is difficult. Despite Marshall's observation 'Our understanding of the brain still lies in the heart of darkness'<sup>2</sup>, some progress has been made. Even Smolensky, after stating 'We simply do not know what architecture the brain uses for performing most cognitive tasks', goes on 'There may be some exceptions, (such as visual and spatial tasks)'<sup>3</sup>.

Historically, there have been two main approaches taken in investigating the operation of the brain.

#### B.1.5.1 The black box approach

Psychologists and others have adopted a "black box" approach. Stimuli in the form of (e.g.) written tests and visual images are submitted to the "black box" (in this case a human subject), the responses noted, and then guesses are made at the mechanisms inside the box that would produce these responses. This approach tells us much about the overall responses of the brain, but not much about its detailed mechanisms of operation.<sup>4</sup>

---

currently no way to specify, obtain approval for supply and obtain delivery, all in the last month at the end of a financial year).

<sup>1</sup>Long, Debra L., Graesser, Arthur C., Long, Charles J., *Four Computational Models for Investigating Neuropsychological Decision-making*, in: *Cognitive Approaches to Neuropsychology*, Williams, J. Michael, and Long, Charles J., Eds., Plenum Press, New York, 1988, p. 22.

<sup>2</sup>Marshall, John C., *Sensation and Semantics*, Nature, Vol. 334 4, Aug 1988, p. 378.

<sup>3</sup>Smolensky, Paul, *On the proper treatment of connectionism*, Behavioral and Brain Sciences, Vol. 11, Cambridge University Press, USA, 1988, p. 1-74.

<sup>4</sup>As an example of this view, see Crick, Francis and Koch, Christof, *Towards a Neurobiological Theory of Consciousness*, CNS Memo 9, January 28, 1991, p. 3 where they assert that this approach is not powerful enough to ever solve a problem or lead to unique answers, but may suggest tentative solutions.

### B.1.5.2 The divide-and-conquer approach

The second approach, used by neuroanatomists and others, is the low-level approach of examining the actual computing elements which make up the brain. Individual computing elements (neurons, see Figure 35) from the brain are examined, and models made of them, (see Figure 36). These models may then may be connected together into a tiny imitation of a brain, (Figures 37, 38). Helmholtz was among first to experiment in this area in 1850,<sup>1</sup> but most work has been done more recently. This approach tells us much about the fine details, but not much about the overall operation of the brain.

Neither approach tells us much about the intermediate-level of brain organisation and operation, although neuroanatomists and artificial life practitioners are starting to attempt to examine this area.<sup>2</sup>

### B.1.5.3 Applying a combined approach

However a combination of the knowledge obtained from the high and low-level approaches, when applied to computer architecture, have resulted in computers referred to as Neural Net machines.

Rosenblatt was one of the pioneers of this approach, he proposed a model of the neuron in 1961, naming it the *perceptron*.<sup>3</sup> He experimented with two-layer networks of perceptrons, (similar to Figure 38, but without the hidden layer). In 1969 Minsky and Papert published a sceptical analysis of Rosenblatt's approach which was so influential that it led to

---

<sup>1</sup>Helmholtz, H. Von, *Preliminary report on the velocity of the nerve impulse*, 1850, reprinted and translated in *Founders of Experimental Psychology*, Blasius, W., Boylan, J., and Kramer, K., Eds., Munich: 25th International Congress of Physiological Science, 1971.

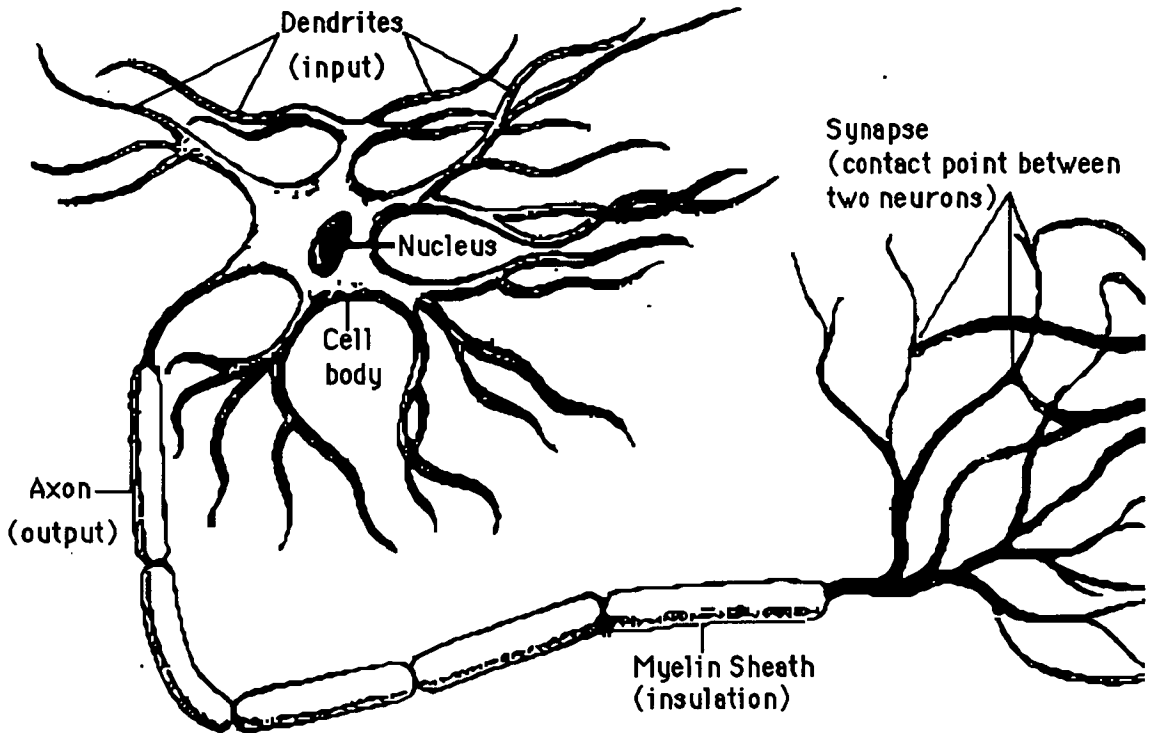
<sup>2</sup>E.g Werner, Gregory M. and Dyer, Michael G., *Evolution of Communication in Artificial Organisms*, Technical Report UCLA-AI-90-06, University of California, Los Angeles, USA, November 1990.

<sup>3</sup>Rosenblatt, Frank, *Principles of Neurodynamics: Perceptrons and the theory of Brain Mechanisms*, Spartan books, Washington, D.C., 1961. Another early neural model was the Adaline (ADaptive LINear Element) of Widrow (1962), for discussion of these see: Zeidenberg pps. 46-51. For a later reference which includes discussion of the mathematical modelling of the function of single neurons, plus applications. see Koch, C., and Segev, I. (Eds.), *Methods in Neural Modelling: From Synapses to Networks*, Bradford Books, MIT Press, Cambridge, USA, 1989.

money for neural network research drying up for nearly 2 decades.<sup>1</sup> However it is notable that Minsky & Papert specifically state in their book that they did not examine three-(or more)-layered networks.<sup>2</sup> It is with this sort of network that Rumelhart<sup>3</sup> and others have produced impressive results, particularly over the last half decade.<sup>4</sup>

## B.2 Modelling the Neuron.

A diagram of a biological neuron is shown in Figure 35.<sup>5</sup>



<sup>1</sup>Originally published in 1969, the MIT press has re-issued an expanded edition including additional notes and corrections, Minsky, Marvin L., and Papert, Seymour A., *Perceptrons*, The MIT Press, Cambridge, Massachusetts, 1988.

<sup>2</sup>*Idem.*, p. 206 of the revised edition.

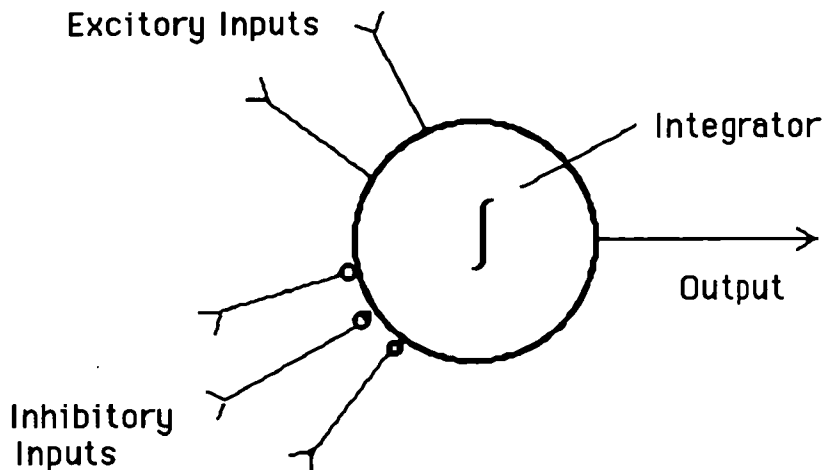
<sup>3</sup>Rumelhart, David E. and McClelland, James L., *Parallel Distributed Processing*, MIT Press, Cambridge, Massachusetts, 1986.

<sup>4</sup>But are they brain models? Mel comments 'while the new generation of more powerful neural-network learning schemes have overcome certain limitations of their single-layer predecessors, they have introduced "anatomical" and "physiological" complexities that, coupled with poor scaling behaviour and the fundamental problem of local minima, make them highly improbable as biological models', see: Mel, Bartlett W., *The Sigma-Pi Column: A Model of Associative Learning in Cerebral Neocortex*, CNS Memo 6, California Institute of Technology, California, 30<sup>th</sup> April 1990, p. 6.

<sup>5</sup>Churchland comments that the brain 'boasts perhaps a hundred distinct and highly specialised cell types, rather than just one'; see: Churchland, Paul M., *A Neurocomputational Perspective*, The MIT Press, Cambridge, Massachusetts, 1992, p. 187. For sketches of some of the varieties of neurons, see Fischbach, p. 29. In neural network simulations, generally only one type is used.

Figure 35 Diagrammatic representation of a Neuron.

Each neuron acts as an individual computing element. There are some 60 different types of neuron known to be used in the human body, many acting as tiny adders, as modelled by the McCulloch and Pitts<sup>1</sup> model of a neuron is shown in Figure 36.<sup>2</sup>

Figure 36 — McCulloch-Pitts model of a Neuron<sup>3</sup>

Each dendrite acts as an input to the neuron. The input may be excitatory or inhibitory.<sup>4</sup> After the level of excitement reaches a

<sup>1</sup>McCulloch, W.S., and Pitts, W., *A Logical Calculus of the Ideas Imminent in Nervous Activity*, Bulletin of Mathematical Biophysics, 5, 1943, pps. 115-133.

<sup>2</sup>Typically a neuron acts as a pulse-code modulator, the frequency being important, (e.g. see Crick and Koch pps. 14-16 where they postulate that neurons forming a feedback loop oscillating at a frequency in the 40-70 Hz range are important in the mechanism of short-term memory and attention). Again, there is speculation that a biological neuron "might well be able to perform several computations at the same time", see: Burrows, Michael and Laurent, Giles "Reflex Circuits and the Control of Movement", in Durbin, Richard, Miall, Christopher and Mitchison, Graeme (Eds.), *The Computing Neuron*, Addison-Wesley, England, 1989, p. 258 ). There are also many other types of specialised neurons, e.g. see the discussion of neurons in the visual system in Barr, Murray L. and Kiernan, John A., *The Human Nervous System*, (fourth edition), Harper and Row, Philadelphia, 1983, p. 301. Thus a "real" neuron is much more complicated than the McCulloch-Pitts model, but this model is a reasonable start; see Anderson, James A., *Cognitive and Psychological Computation with Neural Models*, IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-13, No. 5, September/October, 1983, p. 799.

<sup>3</sup>This is a markedly simplified model of a neuron. Bullock (quoted by Miall, Christopher, "The Diversity of Neuronal Properties", in Durbin, Richard, Miall, Christopher and Mitchison, Graeme (Eds.), *The Computing Neuron*, Addison-Wesley, England, 1989, p. 12 ) presents "a list of 46 separate variable properties of neurons, of which 23 clearly have some temporal dependence, another 7 are activity dependant". Miall goes on to comment on his attempts to model a neuron including some temporal effects, (Miall, pps. 21-31). Koch and Segev also cover this area. However, even though simplified, the McCulloch-Pitts model still produces some very useful results.

<sup>4</sup>Stevens, Charles F., *The Neuron*, in *Progress in Neuroscience*, W.H. Freeman and Company, New York, 1986, p. 5.

certain level, the cell "fires",<sup>1</sup> and a signal passes down the axon, and is distributed to all the axonic connections, which will pass the signal to the dendrites of other neurons via synapses.<sup>2</sup>

In the case of biological systems, the charges are varied by varying the electrolyte level in the cell.<sup>3</sup> In the case of a hardware implementation of a McCulloch-Pitts neuron, an electrical charge may be used. In the case of a software simulation, counters are sufficient.<sup>4</sup>

In the case of biological systems, the charges are mainly transmitted from axon to dendrite across an actual gap by chemical neurotransmitters such as acetylcholine and dopamine.<sup>5</sup> Again, the methods of electronic imitation will be obvious to those with some software experience.

The individual neuron is, by computer terms, an extremely slow device. It responds in milliseconds, (see ref. 2 on page 275 of this thesis). By comparison, switching times of silicon VLSI

---

<sup>1</sup>A swing of about 70 millivolts, negative inside the axon (the potassium equilibrium potential) to about 40 millivolts, positive inside the axon (the sodium equilibrium potential). The swing lasts about 1 millisecond, with approximately a 2 millisecond recovery period; see Groves, Philip and Schlesinger, Kurt, *Biological Psychology*, Wm. C. Brown Company, Dubuque, Iowa, 1979, p. 102; also see a more detailed related discussion in: Keynes, Richard D., *The Nerve Impulses and the Squid*, Physiological Psychology, W.H. Freeman and Company, San Francisco, 1972, p. 128.

<sup>2</sup>The neurotransmitters carry the electrical pulse across a synaptic cleft of about 20 millimicrons. There are some 'direct' electrical connections within the human body (with a 'gap' of about 2 nanometres), but these are rare and synaptic clefts dominate signal transmission. For more details of the synapse, see: Eccles, Sir John, *The Synapse*, Physiological Psychology, W.H. Freeman and Company, San Francisco, 1972, p. 136.

<sup>3</sup>Leibovic, K. Nicholas, *Phototransduction in Vertebrate Rods: An Example of the Interaction of Theory and Experiment in Neuroscience*; IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-13, No. 5, September/October, 1983, pps. 732-741; also Lewis, Edwin R., *The Elements of Single Neurons: A Review*, IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-13, No. 5, September/October, 1983, pps. 702-710.

<sup>4</sup>Whilst the McCulloch-Pitts neuronal model is the most widely used, some authorities have discussed including more of the biophysical properties of single cells in the simulations, plus examining their effect on the dynamics of the operation of networks, e.g. see Koch & Segev.

<sup>5</sup>Thompson, Richard F., *The Brain*, W.H. Freeman & Company, New York, 1985, p. 103 onwards. It is of interest that many of the so-called 'recreational' drugs mimic the effects of the neurotransmitters on certain nerve cells and hence function as mood-altering substances, e.g. the effect of acetylcholine is mimicked at some synapses by nicotine; Groves and Schlesinger, p. 131; ethyl alcohol is one of the most powerful agents known for stimulating the production of dopamine.; also cocaine 'binds to and inhibits a protein that transports dopamine away from its site of action,[and] is one of the most powerful reinforcing drugs known', Fischbach, p. 30. A receptor for marijuana has also been discovered; see also Holloway, Marguerite, 'Rx for addiction', *Scientific American*, Volume 264, Number 3, March 1991, pps. 71-79.



devices produced by this University are in the range of low to fractional nanoseconds, with experimental gallium-arsenide VLSI being some 10-20 times faster. However most present-day VLSI chips have only one processor, the human brain has an estimated  $10^{11}$  neurons.<sup>1</sup>

The number of inter-CPU connections also varies widely. Each neuron 'is connected to as many as 10,000 others'.<sup>2</sup> Typically computer networks have far fewer interconnections. For comparison, a DAP (Distributed Array Processor) has a set of CPUs arranged in a notionally rectangular grid, with the processors connected to the four nearest processors. A cubic array would need 6 connections, and a hypercube array a notional 8 connections. No hardware implementation so far has, to my knowledge, allowed the richness of connections available biologically; for example the transputer only allows direct connection to four neighbouring processors, nowhere near sufficient to implement neural networks directly, as will be seen in the following discussion.<sup>3</sup>

### B.3 Joining the model neurons into a network.

The model neurons may be connected into networks in several different ways. The eight main classifications of neural network architectures are shown in Figure 37.<sup>4</sup>

---

<sup>1</sup>Thompson, Richard F., *Progress in Neuroscience*, W.H. Freeman and Company, New York, 1986, p. 2.

<sup>2</sup>Hecht-Nielsen, p. 37; other authorities suggest an average of about  $10^3$  connections per neuron.

<sup>3</sup>Evans & Deehan (p. 55) reviewing Minsky's work, comment "Nerve cells, after all, do not have intelligence of their own. Yet group 100 thousand million of them together and they do". There has been much experimentation in the ways the mathematically modelled neurons may be connected together, the gleam in the back of the researcher's mind often being that a particular architecture may well allow realistic simulation of that Holy Grail of AI, intelligence.

<sup>4</sup>These are the main types of Neural Networks which have a reasonably established mathematical base. There are also many other topologies employed by biologists attempting to produce computer models of the actual neural processes they observe in animals, sometimes claiming better results than those achieved by the networks listed above, e.g. see: Alkon, Daniel L., *Memory Storage and Neural Systems*, Scientific American, July 1989, pps. 33-34, where claims of markedly reduced learning times are made for the author's DYSTAL (Dynamically Stable Associative Learning) system. This system is also more completely described in: Alkon, D.L., Blackwell, K.T., Barbour, G.S., Rigler, A.K. and Vogl, T.P., *Pattern-Recognition by an Artificial Network Derived from Biologic Neuronal Systems*, Biological Cybernetics, No. 62, 1990, pps. 363-376.

The following sections comment very briefly on the Hopfield (B.3.1), Hamming (B.3.2), BAM (B.3.3), Kohonen (B.3.5) and Neocognition (B.3.6) nets and the Carpenter/Grossberg classifiers (B.3.4). More detail is given about multi-layer perceptron nets in section B.3.7. Specific comments on the 3-layer perceptron net are made in section B.3.7.1, and the ability of the 3-layer perceptron net to generalise from its training data to handle previously unseen data is noted in section B.3.7.2. Section B.3.7.3 notes there are some cases when it can be advantageous to use more than 3 layers of perceptrons.

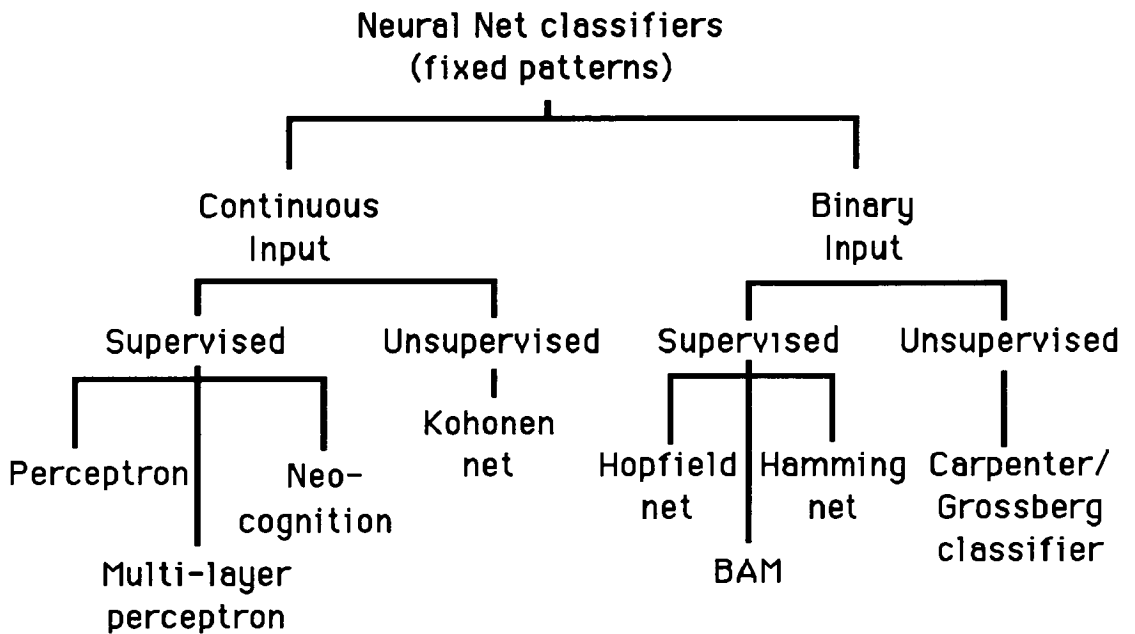


Figure 37 — Main Types of Neural Networks.

Briefly:-<sup>1</sup>

The 'supervised' networks need to be 'trained' by presenting corresponding pairs of sample inputs and the required responses. The 'unsupervised' nets do not need teaching, but use a 'clustering' methodology to group examples which have similar attributes; (this method generally needs more time to settle on a classification).<sup>2</sup> In both cases the nets can then use the knowledge so gained to classify subsequent input.

<sup>1</sup>For a more detailed discussion of the various types of nets, see: Lippmann, Richard L., *An Introduction to Computing with Neural Nets*, IEEE ASSP Magazine, April 1987, pps. 4-27; see also: Pao, Yoh-Han, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1989.

<sup>2</sup>For a theoretical background to unsupervised nets, see: Amari, Shun-Ichi, *Field Theory of Self-Organising Neural Nets*, IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-13, No. 5, September/October 1983, pps. 741-748.

### B.3.1 Hopfield nets.

Hopfield neural nets are suitable for problems involving binary input, e.g. character recognition. The nets have produced impressive results in correctly classifying very distorted images. The disadvantage is that the nets must be large, generally the number of images to be recognised is less than  $0.15N$ , where  $N$  is the number of input signals. To recognise 10 classes might require more than 70 nodes and 5000 connection weights, (for more details of the purpose of the 'weights', see Figures 41 and 43). This type of net is particularly suitable for VLSI implementation as the weights are set in advance.<sup>1</sup> Also Lawrence comments they 'are especially good for finding the best answer out of many possibilities.'<sup>2</sup>

### B.3.2 Hamming nets

The Hamming net may be regarded as an 'optimised' version of the Hopfield net, which at worst has a performance equal to the Hopfield net, and is usually better.<sup>3</sup>

### B.3.3 Bi-directional Associative Memory nets

The BAM (Bi-directional Associative Memory) is the Hopfield network in a generalised form, and, while a trained feedback model is much more complicated than the original Hopfield network, it does take the network to its logical conclusion.<sup>4</sup>

### B.3.4 Carpenter/Grossberg classifiers.

The Carpenter/Grossberg classifier forms clusters, and thus can form classifications without being taught. It has generally less neurons than either of the preceding nets, and can perform well with perfect input patterns, but even a small amount of noise causes problems.<sup>5</sup>

---

<sup>1</sup>Lippmann, p. 7; see also Pao, p. 155.

<sup>2</sup>Lawrence, Jeanette, *Untangling Neural Nets*, Dr. Dobbs Journal, M&T Publishing Inc., Redwood City, California. April 1990, p. 40.

<sup>3</sup>Lippmann, p. 9; see also Pao, p. 174.

<sup>4</sup>Lawrence, p. 42.

<sup>5</sup>Lippmann, p. 11; see also Pao, p. 179, 183.

### B.3.5 Kohonen nets

The Kohonen net also clusters, and can perform better than the Carpenter/Grossberg classifier under conditions of noisy input, but requires much more extensive training, and (because it uses Hebbian training) prefers input patterns which are orthogonal.<sup>1</sup>

### B.3.6 Neocognition nets

The Neocognition net has been proposed in both an unsupervised, and more recently, a supervised form which uses a teacher. This type of network can be dynamically created, saving work by the user.<sup>2</sup> Regarding this topology, Lawrence comments:

The multi-layer (seven- or nine- layer) system assumes that the builder of the network knows roughly what kind of result is wanted. ... It uses a variation of the Hebbian Rule. After learning is complete, the final Neocognition system is capable of recognising handwritten numerals presented in any visual field location, even with considerable distortion. Drawbacks of the Neocognitron are that it is highly specialised and requires a large number of neurons and connections.<sup>3</sup>

### B.3.7 Multi-layer perceptron nets

These nets excited much interest when first proposed 'in the 1960s under the rubric of "perceptrons"'.<sup>4</sup> Two layer nets were shown to be able to perform some simple classifications well. However, as Minsky and Papert pointed out, they cannot handle an exclusive-OR classification.<sup>5</sup>

The multi-layer perceptron<sup>6</sup> net overcame many of the limitations of the perceptron net, (including the ability to handle

---

<sup>1</sup>Lippmann, p. 19; see also Pao, p. 182. This type of net was also proposed by J.A. Anderson independently of Kohonen, (Lawrence, p. 43).

<sup>2</sup>Czuchy, Andrew J., *A Neural Network Instantiation Environment*, Dr. Dobbs Journal, M&T Publishing Inc., Redwood City, California. April 1990, p. 28.

<sup>3</sup>Lawrence, p. 44.

<sup>4</sup>Waldrop, M. Mitchell, *Complexity - the emerging science at the edge of order and chaos*, Penguin books, London, 1994, p. 181.

<sup>5</sup>Lippmann, p. 13; see also Pao, p. 115.

<sup>6</sup>Lippmann p. 15; see also Pao p. 120. Tesauro comments 'the multi-layer architecture, given sufficient hidden units, is capable of universal function approximation', see: Tesauro, Gerald, 'Temporal Difference Learning in Backgammon Strategy', in Sleeman, Derek and Edwards, Peter, *Machine*

an exclusive-OR). They have only recently become practical as effective training algorithms have been devised. Even though there is still some controversy as to whether this architecture accurately imitates brain functioning,<sup>1</sup> it has been used in classification problems such as speech and character recognition and noise filtering in both electrical signals and image processing, and is the best at generalising.<sup>2</sup>

A multi-layer perceptron net was implemented to test its applicability to species classification. This will now be discussed in more detail.

### B.3.7.1 A 3-layer perceptron net

In a 3-level multi-layer perceptron net the neurons are connected as shown in Figure 38.<sup>3</sup>

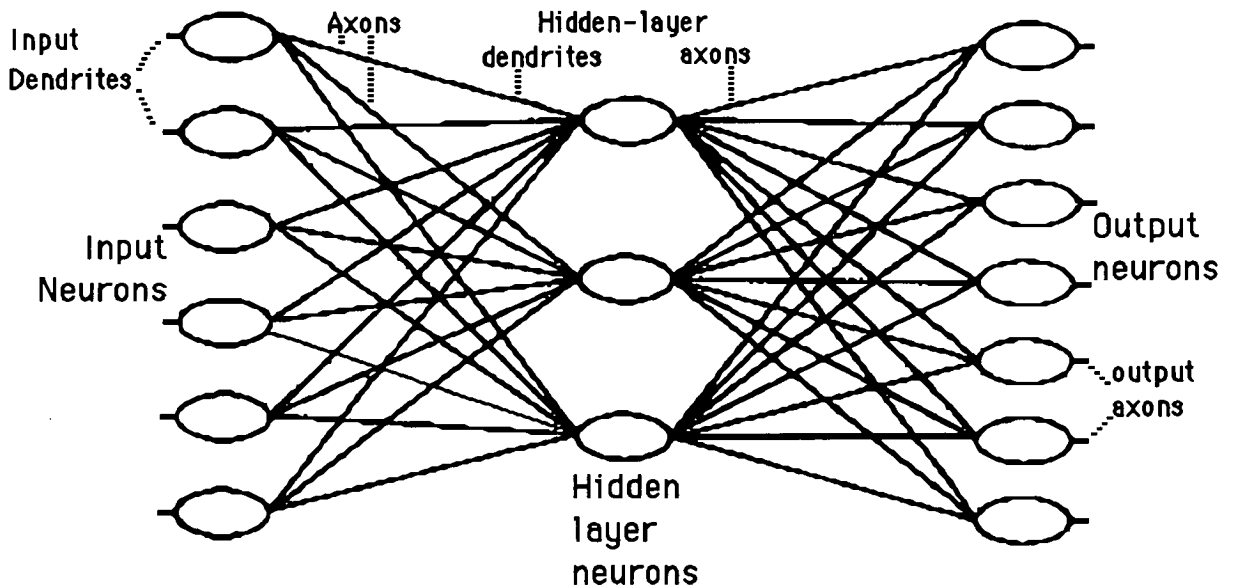



Figure 38 — Three-layer perceptron net

*Learning Proceedings of the Ninth International Workshop*, Morgan Kaufmann Incorporated, 1992, pps. 451-457.

<sup>1</sup>One of the reasons there is still controversy is that it is still uncertain what processes the brain uses and how it uses them! However there are some known differences between perceptron nets and brain functioning, for a useful discussion see: Churchland, Paul M., *A Neurocomputational Perspective*, The MIT Press, Cambridge, Massachusetts, 1992, pps. 181-188.

<sup>2</sup>For a discussion of filtering a noisy signal (an electrocardiogram taken from a patient who was illuminated by fluorescent lights), see: Klimasauskas, Casy, *Neural Nets and Noise Filtering*, Dr. Dobbs Journal of Software Tools, January 1989, p. 32.

<sup>3</sup>Other connections are possible, e.g. Sontag, Eduardo D., *Feedforward Nets for Interpolation and Classification*, SYCON - Centre for Systems and Control, Department of Mathematics, Rutgers University, New Brunswick, examines the effects of directly connecting the input and output layer neurons.

Each  represents one neuron. The input layer of Figure 38 has six neurons, three hidden-layer neurons, and seven output neurons, and hence may be trained to (e.g.) recognise a six state input and show this by signalling on one of the output neuron axons.<sup>1</sup> This type of network has applications in commerce, e.g. banking, where a clerk can feed in details of a loan applicant's financial status, and get a signal from one of three output neurons to 'accept', 'reject', or perhaps 'refer to the manager'. Similar types of applications occur in insurance underwriting and stocks and share trading to name a few. Note that no algorithmic programming is required to feed in rules to the net, this is particularly useful when no rules are known. Neural networks produce results in these areas by being 'trained' with sample examples, and then using the knowledge so gained to classify subsequent input.

As an example, consider a network configured to recognise printed letters. The input to the network can be obtained from a digitised observation of a letter, such as is represented in Figure 39.

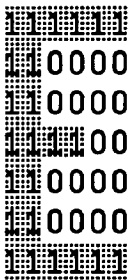


Figure 39 — Binary image of a character

There would be a similar diagram for each letter to be recognised. Each letter would be represented by (in this case) a series of 42 zeros and ones, hence the neural net would be configured with 42 input neurons.

<sup>1</sup>This example network has a total of 16 neurons. Higher numbers may be used, see: Caudill, Maureen, *Neural Network Training Tips and Techniques*, AI Expert, January 1991, p. 58 where she suggests that for a 25 MHz PC-AT 386 with a 387 math co-processor running a three-layer network a maximum of 200-300 total neurons is practical, perhaps double this if a high-speed work station is being used. She also notes that specialised high-speed accelerator cards are available which allow 50-100 times as many connections (*not* total neurons) as would be practical on a PC without the card.

The net would be initially configured with a number of hidden neurons<sup>1</sup> equal to the number of letters to be recognised; 26 if only the upper case alphabet is to be recognised, double that if the lower case letters are also to be learnt by the network, more if punctuation is also to be recognised. (The system may also work satisfactorily with fewer middle layer neurons, but the minimum number cannot be predicted by current theory).

The number of output neurons depends on the type of output signal desired. Suppose the system is configured to distinguish between upper and lower case letters, only two output neurons representing 'upper case' or 'lower case' would be required. If the output required was the ASCII code for the letter, six or seven output neurons would give the ASCII bit pattern. If each input letter were to be represented individually, ignoring case, the net would be configured with 26 output neurons.<sup>2</sup>

### *B.3.7.2 Ability of a 3-layer perceptron net to generalise*

Note that three layer nets of this type are good at recognising regular input, e.g. a fixed font with accurately formed characters. In this case it seems in practice to be useful to set the number of hidden nodes equal to the number of cases to be recognised, as this will often result in faster learning. In the ultimate this could result in hidden layer "grandmother cells" where each hidden cell responds only to one particular input pattern.<sup>3</sup> While this

---

<sup>1</sup>'hidden neurons' can also be referred to as 'middle layer neurons'.

<sup>2</sup>This configuration of network would be similar to the 'winner take all' networks where only one cell is excited, and the rest are inhibited. Anderson notes that Feldman and Ballard, also Barlow, propose this as a suitable mechanism for the brain to detect external stimuli; see: Anderson, p. 800. Restak notes Nobel Prize-winning research that confirms the existence of such cells in the cerebral cortex of cats; Restak, Richard M., *The Brain*, Bantam Books, Toronto, 1984, p. 53. The 'Aristotelian neural net' architecture discussed in section B.6.2.2 of this thesis was developed in this vein.

<sup>3</sup>Also sometimes referred to as "red Volkswagen" cells. In both cases the name is meant to refer to an object which everyone would be able to identify immediately, perhaps with the firing of a single specialised neuron, see: Barlow, H.B., *Single units and sensation, a neuron doctrine for perceptual psychology?*, Perception 1, 1972, pps. 163-169. A comparison can be made with specific-purpose neurons in the human visual system which are "wired" (either epigenetically or experientially) to recognise (fire) only when certain situations are encountered, (e.g. in the case of the visual system, a boundary between light and dark which is at a certain inclination to the vertical, a lighter spot surrounded by a darker area, a specific-length boundary travelling in a certain direction etc., (for further discussion see Hanson & Olsen, pps. 13-23). There is even a suggestion of "Neurons which respond preferentially or selectively to faces..." (several authorities quoted by Rolls, pps. 127-132). Rolls (p. 126) also reports neurons specialised for gustatory stimulation.

approach yields fast learning, it is rarely used as the net can not correctly identify input patterns which have not been previously encountered, i.e. it can not “generalise” from previously learn input.

The ability of neural networks to “generalise” is regarded as important.<sup>1</sup> It is enhanced when the number of hidden nodes is not equal to the number of cases to be recognised, and “noisy” input data is used to train the network.<sup>2</sup> This results in a network which performs better with noisy or novel input which is near (but not identical) to the input data on which the network has been trained. In terms of Figure 44, the “grandmother” case could be regarded as developing a deep and narrow region around the minima for each separate input case. With novel data, cases very near the minima are recognised, but cases in between minima are not. The “generalised” network could be regarded as developing a broader (and, for the same amount of training, usually less deep) region around the minima associated with each input case. The regions near the minima being broader, novel data is more easily tolerated, and identification may be made.

Continued training of the “generalised” network seems to result in deeper, narrower regions around the minima in the “generalised” case as well. Since this results in less ability to classify noisy data, the users who need this generalising ability refer to this type of network as being “overtrained”.<sup>3</sup>

---

<sup>1</sup>Note, however, the previously discussed problems that can arise as a result of inductive inference of this type which is based on incomplete enumeration.

<sup>2</sup>Caudill (p.60) suggests a practical method of finding a useful number of neurons in the middle layer is to start with a low number, and if the network takes too long to train, increase the number of middle nodes by 10%, (and if that doesn't work, try another 10% and so on).

<sup>3</sup>The concept of “overtraining” is similar to the concept of “over-learning” used by Crick and Koch (p. 11) in the case of the human brain, where they use the term to refer to learning which “may be acquired by frequently repeated experience”, and suggests that this type of learning implies that “many of the neurons involved have as a result become strongly connected together”. However they also suggest a mechanism which may assist the memory, attention, (and hence the process of generalisation) in the case of humans. This is the result of the operation of a different type of mechanism than the difference between grandmother and generalised learning noted in the case of neural nets above. They postulate a “spotlight of attention” which “is thought ... to concentrate on one place in the visual field after another, possibly moving every 60 ms or so” which results in a temporary binding between neurons (equivalent to a transitory alteration of weights in a neural net). This temporary binding is postulated to be non-Hebbian, (by comparison with “normal” links which are normally postulated to be Hebbian, i.e. the strength of the links is related to



### B.3.7.3 More than 3 layers in a perceptron net?

More than three layers may be used in the network, however Caudill comments:

be especially careful to stick to three layers unless you have an overriding, truly compelling need to go to four. ... Don't even consider going to five layers. Every time the error is back-propagated to the previous one, it becomes less and less meaningful.

One "truly compelling need" occurs when the user has to deal with input which represents a function which is not continuous, (such as occur in many engineering control applications). Sontag notes that, whilst

a three-layer network is fine for input representing continuous functions, a four layer network is preferable if veridical

---

activity on the particular synaptic junction, Edelman, p. 38, 46; Zeidenberg, pps. 51-54; Linsker, p. 357), and concentrates on one particular observed object at a time, possibly by firing the appropriate recognitional neurons in semi-synchrony (probably in the 40-70 Hz range) thus imposing a temporary unity on these recognitional neurons. This mechanism is postulated by Crick and Koch to correspond to what is called in psychology "short-term memory" (memory which lasts for several seconds, typically containing up to seven items (regarding short-term memory, see also Edelman p. 117)). Most neural nets do not offer any similar facility, only using a mechanism (weights and links which do not change after the initial learning) which could be compared to human long-term memory. This limitation of once-only learning means most neural nets (like un-maintained expert systems) suffer from a similar problem to that suffered by Jimmie G., (see: Sacks, Oliver, *The Man Who Mistook His Wife For His Hat*, Pan Books, London, 1986, pps. 22-41). Jimmie was a healthy and handsome 49 year old who had (and continued to) suffer from a complete loss of short-term memory, remembering nothing that had happened to him since he was 19 years of age; (Korsakov's syndrome, resulting from alcoholism), a problem which was so psychiatrically debilitating that he had to be confined. The message about the relevance of a once-only educated neural net or expert system to a changing environment is clear.

A similar case of memory deficit (with more details of the possible mechanisms involved) is reported by Kandel; see Kandel, Eric R. and Hawkins, Robert D., 'The Biological Basis of Learning and Individuality', *Scientific American*, Vol. 267 No. 3, September 1992, pps. 52-60; and again more comprehensively in McCarthy, Rosaleen and Warrington, Elizabeth K., *Cognitive Neuropsychology A Clinical Introduction*, Academic Press, San Diego, 1990, pps. 275-342 (this latter discussion separates several type of memory in humans, a distinction of significance to computer scientists working in the field of artificial intelligence). It will be noted that the preferable idea of continuous learning, more in the model of human learning, is being pursued, e.g. see: Grefenstette, John J. and Ramsey, Connie Loggia, 'An Approach to Anytime Learning', in Sleeman, Derek and Edwards, Peter, *Machine Learning Proceedings of the Ninth International Workshop*, Morgan Kaufmann Incorporated, 1992, pps. 20-29.

decisions are required for input representing functions containing discontinuities.<sup>1</sup>

Also Fukushima et. al. note that to recognise distorted (e.g. handwritten) characters, a multi-layer architecture is again preferable.<sup>2</sup>

In this context, it is interesting to note that Churchland comments that the brain has 'at least 10 distinct layers of hidden units'.<sup>3</sup> Kuncicky and Kandel make a lower estimate, commenting 'The cerebral cortex operates with approximately six layers of neurons. What implications does this have for PDP networks?'<sup>4</sup>

## B.4 Some Neural Net Theory

Soviet mathematician Andrei Kolmogorov proved the thirteenth problem of Hilbert in 1957, but the significance was not recognised until Robert Hecht-Nielsen restated the theorem for neural nets.<sup>5</sup> This establishes that if we want to map a real vector of dimension M into a real vector of dimension N, this may be done exactly if:

- a) there are exactly M neurons in the input layer
- b) there are exactly N neurons in the output layer
- c) there are at least  $2 * M + 1$  neurons in the middle layer<sup>6</sup>

---

<sup>1</sup>Sontag, Eduardo D., *Feedback stabilization using two-hidden-layer nets*, Report SYNCON-90-11, Rutgers University, New Brunswick, October 1990.

<sup>2</sup>Fukushima, Kunihiko, Miyake, Sei and Ito, Takayuki, *Neocognition: A Neural Network Model for a Mechanism of Visual Pattern Recognition*, IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-13, No. 5, September/October 1983, pps. 826-834. For an alternative architecture, see Fahlman, Scott E. and Lebiere, Christian, *The Cascade-Correlation Learning Architecture*, Report CMU-CS-90-100, Carnegie Mellon University, Pittsburgh, Feb. 1990.

<sup>3</sup>Churchland, Paul M., *A Neurocomputational Perspective*, The MIT Press, Cambridge, Massachusetts, 1992, p. 178.

<sup>4</sup>Kuncicky, David and Kandel, Abraham, 'The weighted fuzzy expected value as an activation function for the parallel distributed processing models', in Zéteñyi, Tamás (Ed.), *Fuzzy Sets in Psychology*, North-Holland, Amsterdam, 1988, p. 229.

<sup>5</sup>Robert Hecht-Nielsen's paper is in the ICNN 87 proceedings, and he extended this work later in the IJCNN 89 proceedings.

<sup>6</sup>Note that, with regard to condition c), the network may also work satisfactorily with less than  $2 * M + 1$  middle layer neurons, but there is so far no theoretically applicable method which will predict how few less will be satisfactory. It is up to the experimenter to achieve an optimum configuration, if this is required.

- d) the transfer functions may be of one variable, are linearly additive, and are non-linearly continuously increasing.

These requirements may be implemented as follows:-<sup>1</sup>

The input signal is fed to an input neuron, (Figure 40), which passes the unmodified signal on to each hidden-layer neuron.

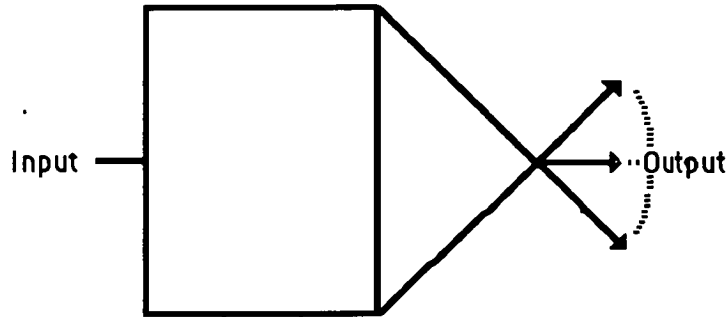


Figure 40 — Diagrammatic representation of an input-layer neuron

The middle-layer neuron (Figure 41) then computes summed input layer  $x$  where:

$$x = \sum W_i * S_i^{\dagger} \text{ where } S_i = \text{signal level of } i^{\text{th}} \text{ dendrite (input).}$$

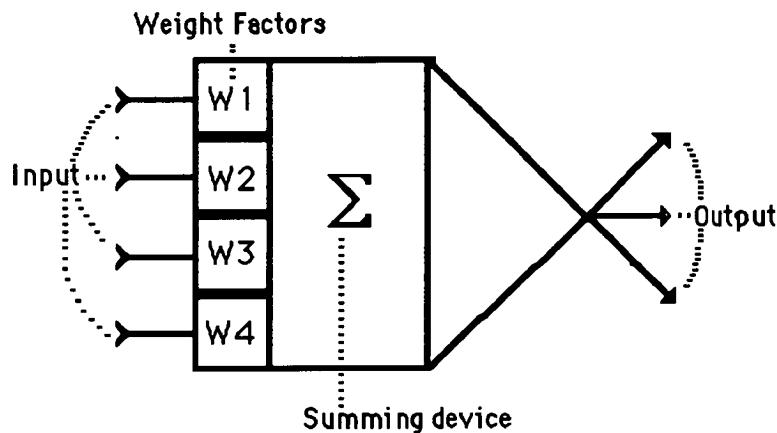


Figure 41 — Diagrammatic representation of a middle-layer neuron

<sup>1</sup>For an interesting introductory article (mainly relating to back-propagation nets) but with some details of other approaches see: Hinton, Geoffrey E., 'How Neural Networks Learn from Experience', *Scientific American*, Vol. 267 No. 3, September 1992, pps. 104-109.

<sup>†</sup> These weights are initially assigned randomly, and are modified by the back-propagation of the error terms, mentioned later in this section. In some specialised cases, however, the weights are more directly representative; e.g. see Hadingham's proposal for a specialised architecture for machine vision: Hadingham, Paul T., *Towards a neural net architecture for rapid learning in machine vision*, Proceedings of the SPIE Conference on Automatic Inspection and High Speed Vision Architectures III, Philadelphia, Pennsylvania, 5-10 November, 1989. (This may also be obtained as Technical Report 89/16, Department of Computer Science, University of Western Australia).

The neuron's activation level  $A$  is then calculated. For this calculation, a non-linear continuously increasing function is needed. The sigmoid function (see Figure 42) is suitable.<sup>1</sup>

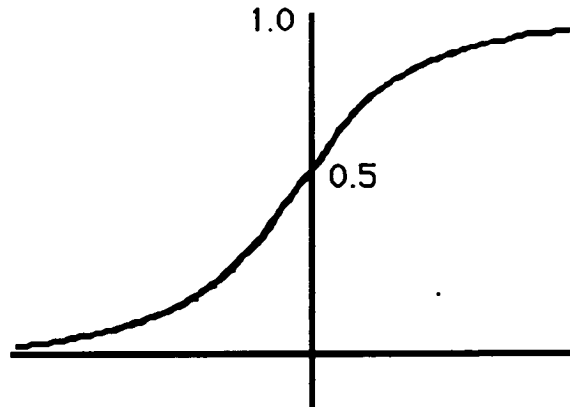


Figure 42 — Shape of a sigmoid function

Thus the activation level  $A$  is:

$$A = \frac{1}{1 + e^{-(x + T)}}$$
 where  $T$  = a threshold value, (often set zero).

This signal level is then distributed to the output neurons (Figure 43).<sup>2</sup>

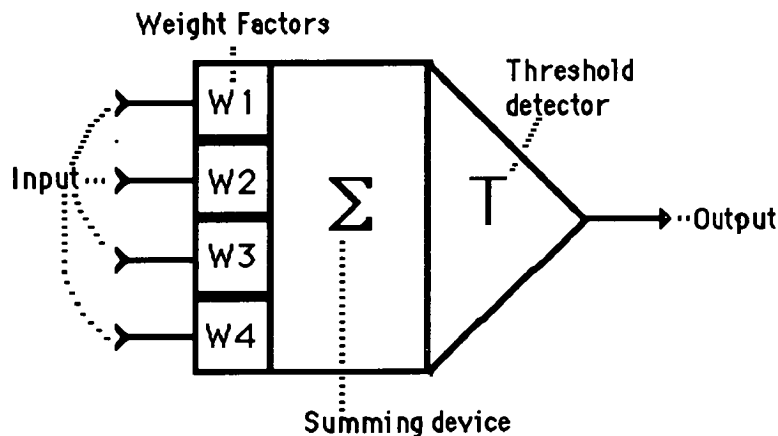


Figure 43 — Diagrammatic representation of an output-layer neuron

<sup>1</sup>Other functions such as tanh are have also been used with good results. Since back-propagation is a gradient descent system that tries to minimise the mean squared error of the system by moving down the gradient of the error curve (Figure 44), it is preferable to use a function that is differentiable, (as in the case of the sigmoid, see Figure 45).

<sup>2</sup>It may be objected that this sum is distributed in an analogue fashion, not as a pulse. There is also an analogue of this process in the human body. Crick and Koch (p. 7) note that the amacrine cells are spikeless. Somjen extends this observation. In the retina, all the pre-ganglionic neurons (horizontal, bipolar and amacrine) operate in an analogue manner. Somjen (p. 125) notes "Whether such impulse-less neurons exist in parts of the CNS other than the retina and olfactory bulb remains to be determined".

Again the activation level is calculated. In this case, if the activation level is greater than a predetermined value (say, 0.95), the neuron signals a "yes", if less than a predetermined value (say 0.05), the neuron does not signal. If the activation level is in between these levels, further learning is needed. This is done by adjusting the output and middle-layer neuron's weights<sup>1</sup> by back-propagating<sup>2</sup> the errors, usually by a steepest-descent method (see Figure 44) such as the delta rule.<sup>3</sup>

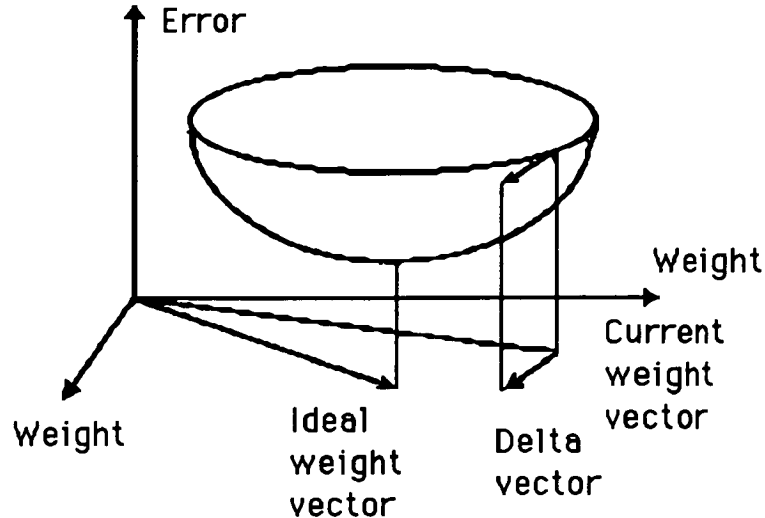


Figure 44 — Delta rule weight changes<sup>4</sup>

<sup>1</sup>The 'weights' in a brain can be adjusted by varying the diameter of the axon (and hence the conduction velocity), and also by varying the placement of synapses, see: Aoki, Chiye and Siekevitz, Phillip, *Plasticity in Brain Development*, Scientific American, December 1988, p. 34; also by varying the permeability of the neural membrane ion channels to potassium-ion flow, see: Alkon, Daniel L., *Memory Storage and Neural Systems*, Scientific American, July 1989, p. 28. Noel Sharkey comments that Donald Hebb (1949) proposed the former mechanism as the way learning and memory occur in the human brain, see: Sharkey, Noel E., *Neural Network Learning Techniques*, in McTear, Michael (Ed.), *Understanding Cognitive Science*, Ellis Horwood Limited, Chichester, 1988, p. 158.

<sup>2</sup>It has been objected that this process is also without parallel in the human neural system. However Somjen (p. 125), referring to the (neural) amacrine cells comments "And, to complicate matters further, the ultra-structure of some synapses suggests two-way or reciprocal function; each of the two participating cells seems to function pre-synaptically at some point, and post-synaptically at another nearby point". This may help explain Barr and Kiernan's (p. 17) initially puzzling reference to the amacrine cell as a neuron which has no axon, only dendrites. A different possible mechanism is proposed by McLaren (in Durbin, Miall and Mitchison (Eds.) pps. 160 - 179) in which an actual feedback circuit is proposed. Similarly Tesauro (pps. 91-101) discusses back-propagation and possible biological neural networks.

<sup>3</sup>There are many methods which have been implemented to speed up this basic back-propagation, often with excellent results, (e.g. Møller's (1990) Scaled Conjugate Gradient Algorithm, Fahlman's (1988) quickprop algorithm and (1990) Cascade-Correlation Learning Architecture), but these will not be discussed in this introductory Appendix.

<sup>4</sup>Unfortunately the error function is often not as smooth as suggested here. In real life, while basically bowl-shaped, the error curve can be a highly complex curve with all sorts of bumps, valleys and hills which contain many local minima. This is the reason random weights are initially used, as it ensures

where:

$$W_{i_{\text{new}}} = W_{i_{\text{old}}} + \frac{C * \text{Error} * P}{|P|^2}$$

where:

C = constant which controls the speed of convergence

Error = difference between observed and desired activation level

P = input pattern vector

|P| = length of input pattern vector.

The derivative of the sigmoid function (see Figure 45) is used when adjusting the weights.



Figure 45 — Shape of derivative of sigmoid function

The cycle of stimuli-presentation, comparison-with-desired-output, and weight-adjustment is repeated until the required comparisons are achieved with no error.<sup>1</sup> This may take (typically) from 24 to 240,000 cycles, depending on the type of delta rule used, and whether the input is bi-valued (boolean) or a continuous variable. Generally continuous input values take longer, whereas binary input values converge more rapidly. Once the "data input/required output" pairs have been learnt, identification occurs after only one forward pass through the network. After the identification program has been loaded into

---

different starting points in the error curve, and hence (with repeated runs) a better chance (but no guarantee) of finding the lowest minima. Random initial weights also help handle the case where convergence is so slow that the network takes longer to converge than the maximum length of run time available to the user, and hence *in practice* "doesn't converge" for all patterns to be learnt. A different starting point may allow convergence within a practical time span. For other hints in helping convergence (varying the weights by other methods, momentum terms, adding noise to input etc.) see: Caudill pps. 56-61.

<sup>1</sup>This is often a time-consuming process. In a specialised application (machine vision) it has been proposed that this may not be necessary, see: Hadingham, paragraph 4.0.

the computer, identification is effectively (as far as the user is concerned) instantaneous on most systems.

Once learnt, the weight values may be retained permanently; however it is interesting to note that, in an attempt to construct neural nets which are more "human", the effects of 'selective ablation' (equivalent to brain damage or forgetting) on accuracy of neural net identification has also been studied.<sup>1</sup>

## B.5 Implementation Issues

Of the types of neural net which could have been implemented, the multi-layer perceptron net was chosen.

Firstly a neural net simulator was developed which adhered to Aristotle's rules of inductive logic, i.e. it assumed complete enumeration, reporting an error if this assumption was found to be invalid. The implementation employed "grandmother cells" of the type discussed in section B.3.7.1 of this thesis, and was developed in Turbo Pascal 4.0 on an IBM-PC clone. The simple data storage methods used in this preliminary implementation imposed a significant limit on the size of net that could be simulated by this implementation.

The use of Aristotelian logic also meant that the neural net simulation was severely restricted in its ability to generalise from the training examples; generalisation could only occur between numerical category limits assuming the result of the observation was a numeric value. This type of generalisation is much more limited in practice than the type of generalisation usually associated with neural nets which implicitly assume inductive logic of the non-Aristotelian type. For this reason the approach was abandoned and development of a second neural net simulation which handled non-Aristotelian inductive logic was started. This development was shelved when the MITRE simulator became available.<sup>2</sup> This was a versatile neural net simulator which ran on a variety of machines. For this work, an implementation running on a Sun 4 was used.

---

<sup>1</sup>Anderson, p. 810.

<sup>2</sup>See Leighton, R., and Wieland, A., *The Aspirin/MIGRAINES Software Tools User's Manual, Release 4.0*, The MITRE Corporation, Washington, 1991.

## B.6 Results Obtained.

Section B.6.1 notes the data treatment necessary to allow the neural nets to be able to use the botanical data. Section B.6.2 details the results obtained.

### B.6.1 Training and Test Data Sets

The data was treated in two ways. Firstly the data was split into training and test sets. Secondly, synthetic data was added.

Computer programs were written which translated data from the standard data format to the format necessary for each of the implementations mentioned above. The programs included the facility to split the data into training and test sets.<sup>1</sup> The specimens were allocated to either of the two sets on an approximate stratified split basis.<sup>2</sup>

A further problem was each implementation's preference for complete data. The approach adopted was that the translation program noted the range of the particular characteristic, and randomly allocated a data value within that range if the data value was missing in the case of that particular specimen. To prevent this "synthetic" data interfering unduly with the training or testing, a facility was added to the translation programs to allow multiple copies of the data to be included in the translated data, the "real" data being the same in each copy of the data, but the "synthetic" data being different random values (each within the noted appropriate range) in each version of the data. The number of multiple versions (the multiplication factor) could be specified.

### B.6.2 Results obtained from Neural Net runs.

Section B.6.2.1 describes the data treatment used. Section B.6.2.2 notes experiences with the Aristotlean net. Section B.6.2.3 presents results obtained from a net specified with the MITRE package.

---

<sup>1</sup>For more details of the programs used to split the data, see section 4.7 b) & g) of this thesis.

<sup>2</sup>See discussion in section 5.4 of this thesis.



### *B.6.2.1 Data Treatment*

Both the *Acaena* and *Danthonia* data were split on an 80%/20% basis. The 80% data was used to train the network. The 20% data was used to test the trained net. The missing values in the *Acaena* data was catered for by use of synthetic data, using a multiplication factor of 20, (see Section B.6.1 for an explanation of this process).

### *B.6.2.2 Experiences with the Aristotlean Net*

The neural net implementation using the Aristotelian assumption of complete enumeration used (almost by definition) categoric input. This prototype trained rapidly.<sup>1</sup> The tested accuracy resulting from the application of this prototype was excellent when applied to pattern-recognition problems similar to character-recognition tasks of the type shown in Figure 39 of this Appendix.

However when applied to the task of botanical species-identification the botanical data was not, in most cases, categoric. Much of the data was in the form of real numbers which had to be categorised before being presented to this form of neural net. This was found to cause several difficulties in practice:-

- a) The accuracy of identification of the botanical species was heavily dependant on the choice of categorisation points.
- b) The categorisation could cause duplicate patterns to occur between species in either or both the test and learning data.
- c) The 80%/20% test data regime could cause a violation of the Aristotlian assumption of complete enumeration of all 20% test patterns in the 80% learning data.

In this investigation the categorisation splitting points were usually chosen on the same basis as those indicated by Selecta-

---

<sup>1</sup>Typically 24 - 50 iterations, compared with the several hundred to several hundred thousand iterations typically required by more usual implementations of neural nets.

key. The results obtained from eight runs of *Danthonia* data are presented in Table 46.

Correctly Classified	Incorrectly Classified	Unable to Classify
52%	44%	4%
47%	41%	12%
46%	50%	4%
45%	48%	7%
45%	40%	15%
49%	45%	6%
48%	45%	7%
51%	44%	5%

Table 46 — Classification Rate, Aristotlean Neural Net method using *Danthonia* Data,

The rate of correct classification is roughly comparable with the rate obtained by some of the statistical methods. However the results were not nearly as good as the results obtained by the Selecta-key methodology. This neural net implementation also recorded much longer training times using the same set of data when compared with the Selecta-key training times. These preliminary results obtained were sufficiently discouraging that work on this software was stopped and work commenced on a prototype which handled real-valued non-Aristotlean data and which hence had a greater ability to generalise.<sup>1</sup>

Development of the second prototype neural net was terminated when the MITRE software tools became available.<sup>2</sup>

<sup>1</sup>Zeidenberg comments 'Without the ability to generalise, neural network models would be like look-up tables, which are not very interesting', see Zeidenberg, Matthew, *Neural Networks in Artificial Intelligence*, Ellis Horwood, New York, 1990, p. 17.

<sup>2</sup>See Leighton, R., and Wieland, A., *The Aspirin/MIGRAINES Software Tools User's Manual, Release 4.0*, The MITRE Corporation, Washington, 1991.

### B.6.2.3 Results Obtained with the MITRE package.

The MITRE package tools permit the specification and construction of neural net simulations. Multi-layer perceptron nets employing a hidden layer were specified. Use of these nets led to the results listed in Tables 47 to 58.<sup>1</sup> In these tables a network configuration of (e.g.) 31-21-11 means a network with 31 input nodes, 21 hidden nodes, and 11 output nodes.

---

<sup>1</sup>The data was split into training and test sets using an 80/20 approximate stratified split. The number of input nodes was set equal to the number of characteristics available in the data. The number of output nodes was set equal to the number of species or taxa to be identified. The number of hidden nodes was varied from the number implied by Kolmogorov's work, (see p. 290 of this thesis) down to a value generally less than the number of output nodes. The nets employed sigmoid transfer functions. Initial runs were on a Sun 3/60, as this was the only machine for which we had the NeWS licence needed by the MIGRAINES interface. The manual noted a marked speed penalty for use of the MIGRAINES interface. Measurements on the Sun 3/60 found this penalty averages about 19% on runs of smaller data. Difficulties with run times led to the transfer of this work to a Sun 4, which produced run times about  $12\frac{1}{2}$  times faster than the Sun 3/60 when each was using only a text-based interface. Even so, run times varied from about 20 minutes to over 4 hours on a lightly loaded Sun 4. The *Danthonia* data converged in about 70,000 to over 4,000,000 iterations; (the lower figure obtained in the case of the nets with the larger number of hidden nodes, the larger figure obtained with the network which had only 8 hidden nodes). In the latter case the learning rate and inertia setting were critical, as one 8 hidden node run which detached when the network went down did not either converge or disastrously diverge after 520,000,000 iterations, before the network came up again and the run was killed. The missing *Acaena* characteristic values also presented a problem. The approach taken was to substitute a random number for each missing value, (the random number being chosen so that it fell within the same range as the range of values observed for that characteristic). The random number generator used produced a rectangular distribution of random numbers. The entire data was then expanded so that it contained 20 copies of the original data, corresponding missing values in each copy of the data being replaced by a different random number in each copy. As might be expected, training was slow, taking from over 2,700,000 iterations for the nets with the smaller number of hidden nodes to less than 275,000 iterations for the larger nets. In all the *Acaena* data runs, a slow training rate with little inertia seemed to be necessary to obtain convergence.

Correctly Classified	Incorrectly Classified	Unable to Classify
53%	12%	35%
63%	11%	26%
57%	14%	29%
67%	7%	26%
55%	17%	28%
60%	14%	26%
55%	7%	38%
62%	12%	26%

Table 47  
Classification Rate, Neural Net method using *Acaena* Data<sup>1</sup>,  
Network configuration = 31 input, 63 hidden and 11 output  
nodes.

Correctly Classified	Incorrectly Classified	Unable to Classify
58%	12%	30%
63%	7%	30%
55%	17%	28%
63%	7%	30%
47%	17%	36%
60%	14%	26%
57%	5%	38%
60%	7%	33%

Table 48  
Classification Rate, Neural Net method using *Acaena* Data<sup>2</sup>,  
Network configuration = 31 input, 41 hidden and 11 output  
nodes.

<sup>1</sup>For details of the data treatment, see section B.6.1 of this Appendix.

<sup>2</sup>For details of the data treatment, see section B.6.1 of this Appendix.

Correctly Classified	Incorrectly Classified	Unable to Classify
56%	12%	32%
63%	21%	16%
57%	17%	26%
63%	7%	30%
45%	19%	36%
52%	10%	38%
55%	7%	38%
55%	7%	38%

Table 49  
Classification Rate, Neural Net method using *Acaena* Data<sup>1</sup>,  
Network configuration = 31 input, 21 hidden and 11 output  
nodes.

Correctly Classified	Incorrectly Classified	Unable to Classify
63%	9%	28%
63%	12%	25%
55%	12%	33%
58%	9%	33%
47%	17%	36%
50%	10%	40%
57%	10%	33%
52%	5%	43%

Table 50  
Classification Rate, Neural Net method using *Acaena* Data<sup>2</sup>,  
Network configuration = 31 input, 11 hidden and 11 output  
nodes.

<sup>1</sup>For details of the data treatment, see section B.6.1 of this Appendix.

<sup>2</sup>For details of the data treatment, see section B.6.1 of this Appendix.

Network Configuration	Correctly Classified	Incorrectly Classified	Unable to Classify
31 - 63 - 11	59.8%	11.8%	28.4%
31 - 41 - 11	57.7%	10.5%	31.8%
31 - 21 - 11	55.8%	11.8%	32.4%
31 - 11 - 11	54.7%	10.5%	34.8%

Table 51  
Average Classification Rate, Neural Net method using *Acaena* Data.

Correctly Classified	Incorrectly Classified	Unable to Classify
47%	6%	47%
62%	5%	33%
53%	2%	45%
72%	5%	23%
60%	6%	34%
43%	2%	55%
54%	7%	39%
56%	4%	40%

Table 52  
Classification Rate, Neural Net method using *Danthonia* Data<sup>1</sup>,  
Network configuration = 41 input, 83 hidden and 19 output nodes.

<sup>1</sup>For details of the data treatment, see section B.6.1 of this Appendix.

Correctly Classified	Incorrectly Classified	Unable to Classify
48%	4%	48%
66%	6%	28%
50%	7%	43%
60%	6%	34%
51%	5%	44%
28%	18%	54%
55%	9%	36%
45%	10%	45%

Table 53  
Classification Rate, Neural Net method using *Danthonia* Data<sup>1</sup>,  
Network configuration = 41 input, 63 hidden and 19 output  
nodes.

Correctly Classified	Incorrectly Classified	Unable to Classify
47%	4%	49%
59%	9%	33%
42%	10%	48%
48%	11%	41%
55%	11%	34%
33%	7%	60%
43%	7%	50%
49%	9%	42%

Table 54  
Classification Rate, Neural Net method using *Danthonia* Data<sup>2</sup>,  
Network configuration = 41 input, 43 hidden and 19 output  
nodes.

<sup>1</sup>For details of the data treatment, see section B.6.1 of this Appendix.

<sup>2</sup>For details of the data treatment, see section B.6.1 of this Appendix.

Correctly Classified	Incorrectly Classified	Unable to Classify
53%	5%	42%
54%	13%	33%
53%	12%	35%
55%	11%	34%
57%	19%	24%
35%	11%	54%
40%	12%	48%
36%	6%	58%

Table 55  
Classification Rate, Neural Net method using *Danthonia* Data<sup>1</sup>,  
Network configuration = 41 input, 23 hidden and 19 output  
nodes.

Correctly Classified	Incorrectly Classified	Unable to Classify
48%	10%	42%
52%	21%	27%
49%	33%	18%
38%	21%	41%
51%	13%	36%
40%	21%	39%
49%	3%	48%
46%	17%	37%

Table 56  
Classification Rate, Neural Net method using *Danthonia* Data<sup>2</sup>,  
Network configuration = 41 input, 13 hidden and 19 output  
nodes.

---

<sup>1</sup>For details of the data treatment, see section B.6.1 of this Appendix.

<sup>2</sup>For details of the data treatment, see section B.6.1 of this Appendix.



Correctly Classified	Incorrectly Classified	Unable to Classify
51%	16%	33%
38%	13%	49%
34%	11%	55%
46%	6%	48%
35%	14%	34%
33%	24%	43%
48%	26%	26%
44%	17%	39%

Table 57  
Classification Rate, Neural Net method using *Danthonia* Data<sup>1</sup>,  
Network configuration = 41 input, 8 hidden and 19 output  
nodes.

Network Configuration	Correctly Classified	Incorrectly Classified	Unable to Classify
41 - 83 - 19	55.8%	4.7%	39.5%
41 - 63 - 19	50.2%	8.2%	41.6%
41 - 43 - 19	46.9%	8.5%	44.6%
41 - 23 - 19	47.8%	11.2%	41.0%
41 - 13 - 19	46.8%	17.2%	36.0%
41 - 8 - 19	42.0%	16.2%	41.8%

Table 58 — Average Classification Rate, Neural Net method using  
*Danthonia* Data.<sup>2</sup>

<sup>1</sup>For details of the data treatment, see section B.6.1 of this Appendix.

<sup>2</sup>For details of the data treatment, see section B.6.1 of this Appendix.

## B.7 Discussion

The classification rates obtained by use of multi-layer perceptron nets are well above the rates which could be obtained on average by chance.<sup>1</sup>

Interestingly, the use of synthetic data did not appear to inhibit the accuracy of the methodology. The identification rate for the *Acaena* data was actually slightly higher than the identification rate observed for the *Danthonia* data, although the chance rate of identification of the *Acaena* data is also higher, and hence the *Danthonia* identification problem is more difficult.

The classification rates for both sets of data were above the rates obtained by use of Clustering methodologies, about the same as those obtained by the Voting methodology, but below those obtained by ID3, Selecta-key and use of Collier's key.

The multi-layer perceptron net had the advantage that no specialist botanical knowledge is needed to train and use the net.

The main disadvantage is that the recognition rate was lower, and the training slower than some of the other methods, (e.g. Selecta-key and Voting).

The learnt data is also in a form which does not allow the extraction of information to form a paper key which could be used for identification of specimens in the field, where a computer is generally not available.

## B.8 Summary

The multi-layer perceptron net gave a rate of identification much better than that which would have been obtained on average by chance.

---

<sup>1</sup> $5\frac{1}{4}$  % in the case of the *Danthonia* data; 9% in the case of the *Acaena* data, if the data contained the same number of specimens per species in each data set. However this was not the case in either of these sets of data. If the user had had a knowledge of the number of specimens identified as belonging to each species, the user could have 'guessed' the percentage of specimens belonging to the largest group of species. If this had been the case, the user could have guessed 9.6% for the *Danthonia* data, 23% for the *Acaena* data.

The classification rates obtained were above those obtained by Clustering methodologies, about the same as those obtained by the Voting methodology, but below those obtained by ID3, Selecta-key and use of Collier's key.

The time taken to train the net was more than the time taken by some of the alternative methodologies (e.g. Voting, Selecta-key) to learn the training data.

A key could not be produced from the learnt data.

# Appendix C: Voting Methods

This Appendix examines a voting methodology proposed for species identification.

In this methodology each characteristic of each specimen of the test data is examined in the light of standards established by training data. Votes are allocated to the species appropriate to each characteristic. The species with the most votes is taken as the most likely species for identification purposes.

Section C.1 of this Appendix discusses the proposed voting methodology. Section C.2 examines results obtained by this methodology. Section C.3 discusses the advantages and disadvantages of this methodology. Section C.4 summarises the use of this methodology.

## C.1 Voting Methodology

The voting methodology was proposed as an offshoot of the Selecta-key methodology. Section C.1.1 discusses the details of the methodology. Section C.1.2 discusses the implementation of the methodology.

### C.1.1 Detail of Methodology

The methodology employs one set of data for training. The results of the training may then be employed to identify specimens.

It is important that the training data should be a statistically valid representative sample of data for the species being considered.

In the Voting methodology each characteristic of the training data is examined separately. A mean and standard deviation for each species group within each characteristic is calculated. For each characteristic, splitting points are established between each species, using a methodology similar to that used by Selecta-key. For each characteristic, the order of species may be different. The order of species and the location of the splitting

points between them are then recorded for each characteristic of the training data.

The results of these calculations may then be used to classify specimens. A test specimen may be classified by comparing each of its characteristic measurements with the splitting points of the corresponding characteristic in the training data. A "vote" is recorded in favour of the species in whose measurement range the test data's characteristic measurement falls. The total number of votes distributed is equal to the number of characteristics measured. The species with the greatest number of accumulated votes is declared to be the likely species to which the test specimen belongs.

### C.1.2 Implementation

The voting methodology was implemented as a computer program running a Sun computer. The program was written in Sun Pascal 2 using the transportable Pascal package developed as part of this project.

When calculating the splitting points within each characteristic's measurements, the program made allowance for those cases where different species had identical means. If the splitting points could not distinguish between species, the vote was divided equally between those species; e.g. if two species could not be distinguished, each of those two species would get half a vote, if three could not be distinguished, each would be allocated one-third of a vote.

After completion of voting, the voting totals for each species were listed. This allowed the user to choose the most likely species. The user could also see if there was a clear "favourite", or if two or more species were very close.

As implemented, the program has the limitation that it only handles cases where complete data is available.<sup>1</sup> It could be extended to handle cases of test data which do not have complete data. Whilst theoretically it could also be extended to

---

<sup>1</sup>Note that this meant this methodology could not be tested with the *Acaena* data, as about three-quarters of the specimens in this data have some characteristic measurements missing.

handle cases of incomplete training data, care should be taken to ensure the implementation recognises cases where the training data is so incomplete that reliable identification of some species may not be able to be undertaken.

## C.2 Results from Voting Methodology

In section C.2.1 the treatment of the data used is discussed. The results obtained using this data are presented in section C.2.2. The results are discussed in section C.2.3.

### C.2.1 Treatment of Data

A previously written computer program was used to split the data into training and test sets.<sup>1</sup>

No individual specimen appeared in both the data sets, but all specimens appeared in one of the two data sets. The specimens were allocated to either of the two sets on an approximate stratified split basis.

### C.2.2 Results from Data

The *Danthonia* data was split on an 80%/20% basis, as outlined in the previous section. The 80% data was used for training. The 20% data was used to test the methodology.<sup>2</sup>

To even out any chance variations which might occur in the random allocations, this process was repeated eight times, and the result totalled. The individual results are shown in Table 59, and the totalled results in Table 60.

---

<sup>1</sup>For more details of this program, see section 4.7 b).

<sup>2</sup>For more details, see section 5.4 of this thesis.

Correctly Classified	Incorrectly Classified
52%	48%
55%	45%
49%	51%
50%	50%
45%	55%
60%	40%
44%	56%
40%	60%

Table 59  
Classification rate — First Choice  
Voting methodology using *Danthonia* data.

The total classification rate is shown in Table 60.

Correctly Classified	Incorrectly Classified
49.3%	50.7%

Table 60  
Total classification rate — First Choice  
Voting methodology using *Danthonia* data.

It was noted that in many cases the first and second choices were very close, and as an indication of this the number of times the correct identification occurs within the first two choices was extracted and is shown in Table 61. The results of eight runs are given, the results being listed in the same order as in Table 59.

Correctly Classified	Incorrectly Classified
67%	33%
65%	35%
59%	41%
65%	35%
55%	45%
72%	28%
64%	36%
63%	37%

Table 61  
First two choices — Voting methodology with *Danthonia* data.

The number of times the correct identification occurs within the first two choices, totalled across the eight runs, is shown in Table 62.

Correctly Classified	Incorrectly Classified
63.8%	36.2%

Table 62  
Total classification rate (first two choices)  
Voting methodology using *Danthonia* data.

C.3 Discussion of Methodology

The Voting methodology was conceived as a quick, simple variation of the Selecta-key methodology. It produced results which were comparable with some of the other methodologies. Section C.3.1 below discusses the results obtained by this methodology, while section C.3.2 discusses the advantages and disadvantages of this method.



### C.3.1 Discussion of Results

In all Tables the identification rate is well above the chance identification rate<sup>1</sup>.

However the recognition rate attained was below that achieved by some of the better methodologies discussed in the main body of this thesis. There are several possible reasons that may contribute to the lower recognition rates.

The Voting methodology, like some of the clustering methodologies, allocates equal importance to all characteristics. If some of the variations between characteristics are not correlated to the species variation, this could lead to less overall discrimination than methods (e.g. Selecta-key) which allow the variation between the discriminatory powers of the various characteristics to be taken into account when identifying specimens.

The Voting methodology would be expected to work best where the data is well separated, and would be expected to suffer proportionately more than other methods (e.g. Selecta-key) when the data is poorly separated. The results obtained appeared to bear out this expectation. As an example, the specimens of *Danthonia pilosa* were much less well separated from their neighbours than the specimens of *Danthonia semiannularis*, and this was reflected in the respective recognition rates. Despite roughly similar average standard deviations, the recognition rates of these two species (extracted from the results which are summarised in Table 59) were 10.2% and 81.0% respectively; (the "first two" recognition rates extracted from Table C.3 results were 22.4% and 92.9% respectively). The results suggest the methodology may well be worth trying if the items to be identified are statistically well separated.

---

<sup>1</sup>  $5\frac{1}{4}$  % in the case of the *Danthonia* data; if the data contained the same number of specimens per species. However this was not the case. If the user had had a knowledge of the number of specimens identified as belonging to each species, the user could have 'guessed' the percentage of specimens belonging to the largest group of species. If this had been the case, the user could have guessed 9.6% for the *Danthonia* data,

### C.3.2 Advantages and Disadvantages

The voting methodology had the appeal of being a simple, quick and relatively straight-forward method. The mean and standard deviation calculations used in the training process could be made using readily available, efficiently coded, standard procedures. After learning, the splitting points appear in a sorted order and an efficient binary chop search algorithm may be used to find the appropriate species to which an allocation of a vote is appropriate. The identification of specimens is thus not much more complicated than a table look-up for each characteristic, and was therefore arguably the computationally most efficient identification algorithm examined in this thesis.

However the method did have the disadvantage of being a fully automatic process that did not offer the advantage that Selecta-key offered of using the expert's preferences. Since these preferences could take account of the expert's knowledge of the limitations of the data being used, it was relatively more important for the success of the voting methodology that the data used was the result of a statistically valid and carefully controlled data collection process.<sup>1</sup>

Given the above restrictions, the voting methodology could be expected to work reasonably well with data which is well separated. The accuracy would be expected to drop sharply when the data consists of poorly separated data, as (unlike Bayes) votes are allocated to only one species in any particular measurement range of each characteristic. The lower recognition rates in the case of poorly separated distributions was confirmed in the case of the data examined.

Despite these disadvantages, the voting methodology produced rates of identification markedly superior to that

---

<sup>1</sup>As mentioned in the main body of this thesis, this may be difficult. Whereas in the case of industrially derived data it is usually a reasonably straight-forward process to plan and achieve comprehensive and representative data collection., the collection of botanic data that is similarly representative of a botanic species is typically much more difficult. The specimens to be measured may be geographically remote, have a geographic distribution that is uncertain, and depend on climatic and soil fertility factors to an unknown extent. For these reasons the extent to which the specimens being measured are typical of the species is often uncertain, and the expert's opinion regarding the relative reliability of various portions of the data is germane. The Voting methodology does not take this into account.

achievable on average by chance. The recognition rate was better than the rate of identification obtained by the various clustering methodologies, in the same range as the neural net results, but below that obtained by the application of Collier's key, ID3 and Selecta-key.

The advantage of the methodology would appear to be that it is a computationally fast and useful methodology of reasonable accuracy suitable to be employed in the classification of statistically well separated botanical species where computer time is at a premium. The results of the learning could also be distributed in a paper format, like a key. In this form it may have some advantages over a key when only partial information about a specimen is available, in that identification would not depend on certain key characteristics of the specimen to be identified being present. However if full information is available, a rich paper key such as Collier's key could be expected to provide a superior rate of identification.

## C.4 Summary

The advantage of the Voting methodology would appear to be that it is a computationally fast and useful methodology of reasonable accuracy. The recognition rate achieved was better than the rate of identification obtained by the various clustering methodologies, in the same range as the neural net results, and below that obtained by the application of Collier's key, ID3 and Selecta-key.

The methodology could be expected to work best when the data is well separated. It will suffer proportionately more than other methods (e.g. Selecta-key) when the data is poorly separated.

The results of the learning could also be distributed in a paper format, like a key. In this form it may have some advantages over a key when only partial information about a specimen is available, in that identification would not depend on certain key characteristics of the specimen to be identified being present. However if full information is available, a paper key such as Collier's key could be expected to provide a superior rate of identification.

# Appendix D:

## Discriminant Analyses

This Appendix examines the results produced by two statistical methodologies used for discriminant analysis.

It should be noted that these methodologies are different from the clustering methodologies (which seek to establish specimens into groups) in that each specimen in the data sets employed must have already been classified before these methodologies can be applied.

During the use of these methodologies each characteristic of each specimen of a test data is examined in the light of standards previously established by use of a training data set. One methodology employed assumes the distribution within each characteristic group to be parametric, the other does not need the assumption that the distributions are parametric.

Section D.1 of this Appendix discusses the proposed methodologies. Section D.2 examines results obtained by these methodologies. Section D.3 discusses the advantages and disadvantages of these methodologies. Section D.4 summarises issues related to the use of these methodologies.

### D.1 Discussion of the Methodologies employed for the Discriminant Analyses

The two types of analyses employed used methodologies proposed by statisticians as being appropriate for the task of discriminant analysis. Section D.1.1 discusses the details of the methodologies. Section D.1.2 discusses the implementation of the methodologies.

#### D.1.1 Detail of Methodologies

Both methodologies employed one set of data for training. It is important that this set of training data should be a statistically valid representative sample of data for the species being considered. The results of the training were then employed to identify specimens in the test set. In this case all specimens in

the test data set had already been classified into species or taxa, and the runs were used to provide an estimate of the accuracy of the methodology. The same methodology could be used to identify specimens whose classification was unknown.

The first methodology used in this Appendix depends on the assumption that each group of observations per characteristic per species can be assumed to be multivariate normal.<sup>1</sup> A parametric method based on multivariate normal distribution theory is then used to derive a quadratic discriminant function.<sup>2</sup> This derived discriminant function, also known as a classification criterion, is then applied to the test data to obtain a classification of the specimens in the set of test data.

The second methodology employed used a non-parametric methodology which does not need the assumption that each group of observations per characteristic per species can be assumed to belong to a normal distribution. Epanechnikov's kernel method was used to generate a non-parametric density estimate in each group in the training data, and to produce a classification criteria which was then applied to classify the test data.<sup>3</sup>

### D.1.2 Implementation

Both methodologies were implemented in release 6.07 of the SAS statistical package, running on a Sun 4.

## D.2 Results obtained by applying Discriminant Analysis Methodologies

In section D.2.1 the treatment of the data used is discussed. The results obtained using this data are presented in section D.2.2. The results are discussed in section D.2.3.

---

<sup>1</sup>This may be a rash assumption in the case of some of the groups of data; see Appendix E, section E.4.1, where the assumption of normality is examined. Since some of the groups do not appear to fit the assumption of multivariate normal distributions, the results of the application of this discriminate analysis will be effected by the robustness of the test to the presence of non-normal groups of data.

<sup>2</sup>For further detail, see: - SAS Institute Inc, *SAS/STAT User's Guide, Release 6.03*, SAS Institute Inc., Cary, NC, 1988, pps. 360 - 363.

<sup>3</sup>For further information, see: SAS Institute Inc, *SAS/STAT User's Guide, Release 6.03*, SAS Institute Inc., Cary, NC, 1988, pps. 363 - 366

### D.2.1 Treatment of Data

Previously written computer programs were used to split the data into training and test sets.<sup>1</sup>

An approximate stratified split methodology was used to split the data into training and test sets. The data translation program also produced a batch file to facilitate running the SAS program.

Using data and batch files produced by these processes, the implementation ran satisfactorily and provided results that could be used for comparison with the Selecta-key methodology.

### D.2.2 Results from Data

Both the *Danthonia* and *Acaena* data were split on an 80%/20% basis, as outlined in the previous section. The 80% data was used for training. The 20% data was used to test the methodology.<sup>2</sup>

Section D.2.2.1 presents results obtained by use of this data with the methodology which assumes multivariate normal distributions, and section D.2.2.2 the results obtained by the methodology which makes no normal assumptions.

#### *D.2.2.1 Results from a parametric methodology*

To even out any chance variations which might occur in the random allocations, the process which obtained the 80%/20% data split was repeated eight times with the *Danthonia* data. Each of the eight data sets was presented to the test which assumed multivariate normal distributions, and the results listed in Table 63 were obtained.

---

<sup>1</sup>For more details see section 4.7 b) & g) of this thesis.

<sup>2</sup>For more details, see section 5.4 of this thesis.

Correctly Classified	Incorrectly Classified
66%	34%
55%	45%
62%	38%
65%	35%
51%	49%
51%	49%
48%	52%
53%	47%

Table 63  
Classification rate — multivariate normal methodology using *Danthonia* data.

The average *Danthonia* classification rate is shown in Table 64.

Correctly Classified	Incorrectly Classified
56%	44%

Table 64  
Average classification rate — multivariate normal methodology using *Danthonia* data.

Similarly, when considering the *Acaena* data, the 80%/20% data allocation split was repeated seven times. Each of the data sets was presented to the test which assumed multivariate normal distributions. The SAS restriction requiring complete data for each specimen eliminated approximately three-quarters of the *Acaena* data from consideration. This meant that the total number of specimens remaining in the test data was small, and for this reason individual results from each run are not presented. The remaining completely described specimens correctly and incorrectly identified were summed across all seven runs, and the overall result obtained presented in Table 65.

Correctly Classified	Incorrectly Classified
43%	57%

Table 65  
Total classification rate for completely described specimens —  
multivariate normal methodology using *Acaena* data.

It should be stressed that the rate of identification in Table 65 is the rate obtained for completely described specimens only. If all test data specimens (both completely and incompletely described) are included, the rate of identification would be much lower. For the record, the rate obtained is as shown in Table 66.

Correctly Classified	Incorrectly Classified	Unable to Classify
9%	12%	79%

Table 66  
Total classification rate for all specimens — multivariate normal  
methodology using *Acaena* data.

*D.2.2.2 Results from a non-parametric methodology*

To even out any chance variations which might occur in the random allocations, the process which obtained the 80%/20% data split was again repeated eight times with the *Danthonia* data. Each of the eight data sets was presented to the test which did not need to assume multivariate normal distributions, and which instead used Epanechnikov's kernel method to generate a non-parametric density estimate in each group in the 80% training data to produce a classification criteria which was then applied to classify the 20% test data. The results listed in Table 67 were obtained.



Correctly Classified	Incorrectly Classified	Unable to Classify
68%	29%	3%
76%	23%	1%
74%	24%	2%
75%	23%	1%
78%	21%	1%
79%	21%	0%
65%	35%	0%
75%	24%	1%

Table 67  
Classification rate — Epanechnikov's kernel methodology using *Danthonia* data.

The average *Danthonia* classification rate is shown in Table 68.

Correctly Classified	Incorrectly Classified	Unable to Classify
74%	25%	1%

Table 68  
Average classification rate — Epanechnikov's kernel methodology using *Danthonia* data.

Again, the 80%/20% data allocation split was repeated seven times using the *Acaena* data. Each of the data sets was presented to the test which used Epanechnikov's kernel method to generate a non-parametric density estimate in each group in the 80% training data to produce a classification criteria which was then applied to classify the 20% test data. The SAS restriction requiring complete data for each specimen again eliminated approximately three-quarters of the *Acaena* data from consideration. This again meant that the total number of specimens remaining in the test data was small. The Epanechnikov's kernel methodology responded to these limitations by producing identification rates which were not

nearly as good as those obtained in Tables 65 and 66, and for this reason are not included here.

### D.2.3 Discussion of Results

In the case of the *Acaena* data, the SAS requirement for complete data for each specimen made the results obtained from each of the methodologies used less than satisfactory. Other implementations of the methodologies which allowed use of the partial data available may have produced better results, but in the absence of these implementations a firm opinion can not be offered. This restriction requiring complete data to be supplied in the case of all specimens in the data set can be a severe restriction in the case of botanical data, and probably eliminates these methodologies as the methodologies of choice if the data to be examined contains a significant number of incompletely described specimens.

In the case of the (complete) *Danthonia* data, the identification rate obtained was well above the chance identification rate for both methodologies.<sup>1</sup>

The identification rate obtained for the discriminant analysis methodology employing the assumption of multivariate normal data (Tables 63 and 64) was above the rates obtained by the application of clustering methodologies,<sup>2</sup> and similar to that obtained by neural net methodologies,<sup>3</sup> but below that obtained several of the other methodologies.

The identification rate obtained by use of Epanechnikov's kernel methodology with the *Danthonia* data was very good, exceeding most of the other methodologies, with the exception of Selecta-key.<sup>4</sup>

---

<sup>1</sup> $5\frac{1}{4}$  % in the case of the *Danthonia* data; if the data contained the same number of specimens per species in each data set. However this was not the case in either of these sets of data. If the user had had a knowledge of the number of specimens identified as belonging to each species, the user could have 'guessed' the percentage of specimens belonging to the largest group of species. If this had been the case, the user could have guessed 9.6% for the *Danthonia* data.

<sup>2</sup>For further detail, see Appendix A, or summary Tables 20 and 21 in the main body of this thesis.

<sup>3</sup>For further details see Appendix B, or summary Tables 22, 23 and 24 in the main body of this thesis.

<sup>4</sup>This is consistent with Ripley's comment that 'Comparisons with other methods are rare, but when done carefully often show that statistical methods

## D.3 Advantages and Disadvantages

Both methodologies had the appeal of being available as a standard part of a readily available commercial package.

Both methodologies had the disadvantage of being fragile in the presence of incompletely described specimens, ignoring these incompletely described specimens completely.

The methodology which assumed multivariate normal distributions appeared not to be as well suited by the data as the approach employing Epanechnikov's kernel methodology. The latter had the advantage of producing identification rates superior to all other methods with the exception of Selecta-key. Compared with Selecta-key, it had the disadvantage of requiring a computer to produce the identification, whereas Selecta-key produced a more readily transportable paper key.

## D.4 Summary

Two statistical discriminant analysis methodologies are introduced. One is a parametric methodology assuming multivariate normal distributions in the groups of data. The second is a non-parametric approach employing Epanechnikov's kernel method of density estimation.

Both methodologies produced good identification rates.

The parametric methodology produced identification rates superior to those obtained by clustering and neural net methodologies.

The non-parametric approach produced identification rates superior to all methodologies with the exception of Selecta-key.

Both statistical approaches have the disadvantage that they need a computer to allow identification, whereas the Selecta-key methodology produces its identifications from a more readily transportable paper key.

---

can out-perform state-of-the-art neural networks'; Ripley, B. D., 'Statistical Aspects of Neural Networks', invited lecture for SemStat (Séminaire Européen de Statistique), Sandbjerg, Denmark, 25<sup>th</sup>-30<sup>th</sup> April 1992. To appear in the proceedings to be published by Chapman & Hall in January 1993, the quotation is from p. 2 of the pre-print.

# Appendix E: Data Used

It was regarded as important that the data used for comparative runs between the methodologies represent the problems typically found in botanic data.<sup>1</sup> This Appendix presents the results of checks carried out on the data selected for use in the comparative runs used in this thesis.

Section E.1 of this Appendix notes the origin of the data used in the tests employed in this thesis. Section E.2 comments on the type of data used, and the reason these sets of data were chosen. Section E.3 comments on the suitability of the data characteristics chosen to specify the species being examined. Section E.4 comments on the likely consistency of the data. Section E.5 summarises the results of this examination of the data.

## E.1 Origins of the Data

The *Acaena* and *Danthonia* data are the main sets of data used in the tests employed in this thesis.

The *Acaena* data was provided by Dr. Tony Orchard.<sup>2</sup>

The *Danthonia* data was provided by Mr. Phil Collier.<sup>3</sup>

The author wishes to thank both Dr. Orchard and Mr. Collier for their generosity in making this data available.

## E.2 Types of Data used.

The *Acaena* data consists of a set of categoric observations and numeric measurements made of characteristics of specimens of 11 taxa of the *Acaena ovina* complex gathered from South-Eastern Australia.<sup>4</sup> The classifications of the *Acaena* taxa

---

<sup>1</sup>These requirements are discussed in sections 2.2.3 and 5.2 of this thesis.

<sup>2</sup>Dr. Orchard is Curator of the Tasmanian Herbarium, and is recognised as a distinguished expert in his field.

<sup>3</sup>Mr. Collier is a Senior Lecturer on the staff of the University of Tasmania. He has several publications in the area of identification of Tasmanian native species, and is recognised to be a leading expert in Tasmania in this field.

<sup>4</sup>A map showing the world-wide distribution of the genus *Acaena* is to be found in: Humphries, Christopher J. and Parenti, Lynne R., *Cladistic Biogeography*, Clarendon Press, Oxford, 1989, Figure 1.5, p. 6.

were made by Dr. Orchard, who also chose the characteristics to be measured, and made the measurements concerned.

The *Danthonia* data consists of a set of categoric observations and numeric measurements made of characteristics of specimens of the 19 Tasmanian species of the *Danthonia* genus. The specimens were classified by Mr. D.I. Morris, who also kindly made them available for further study.<sup>1</sup> The choice of characteristics to be measured was made by Mr. Morris. The measurements of the *Danthonia* characteristics were made by technical staff employed for this purpose. Some of the characteristics were difficult to observe (having to be examined via a microscope), and as a consequence, the resulting measurements may be of a lower standard than the rest of the data. It should be noted that the measurements and observations were obtained subsequent to the classification of the data, and that the classifications did not depend upon these measurements.

These data sets were chosen as they are good representative samples of the type of data obtainable from botanic sources, and contain examples of the type of problems which are common in data of this sort, e.g. missing data. In many other settings, missing data is not such a great problem, as another set of measurements can often be obtained fairly readily. This may be much more difficult in the case of botanic data, as suitable specimens may be geographically remote, in an inappropriate stage of development (e.g. not flowering) and in cases such as *Acaena echinata* var. *robusta*, possibly very hard to locate, (this taxa has not been observed for many years and is possibly extinct). It is thus relatively more important that inductive learning methods used with botanic data must, in practice, be capable of more flexibility in dealing with problems of this type than methods used with other (e.g. industrial) data. It is for this reason that realistic data has been chosen.

The data chosen also have the advantage that they have been used in previous work. This facilitates comparison between other methods and the methods used in this thesis.

---

<sup>1</sup>Mr. D.I. Morris is an Honorary Botanist at the Tasmanian Herbarium who has recently completed work on the taxonomy of Tasmanian grasses for publication.

## E.3 Suitability of the Data Characteristics used.

Gammack quotes Kidd:

... it is vital that ... there is a high degree of cognitive compatibility between user and system. It must employ similar knowledge structures.<sup>1</sup>

In the case of data of botanical origin, this suggests as a minimum that the knowledge envelope, (that is, the characteristics measured), has to be conformant with those employed by competent botanists. The characteristics used for both the *Acaena* and *Danthonia* data met this requirement, as for both data the characteristics to be observed were specified by an acknowledged expert in this field.

However there are also statistical requirements to be met. It would often be useful if all characteristics measured were completely independent.<sup>2</sup>

Tables 69 and 70 list the mutual correlations of the *Acaena* and *Danthonia* data characteristics, averaged across all taxa. In practice it could be suggested that a correlation within the range  $\pm 0.2$  would be adequate to meet the requirement of independence. Correlations outside this range, but within the range  $\pm 0.4$  would suggest the likelihood of a small relationship between the characteristics; from  $\pm 0.4$  to  $\pm 0.7$  a moderate relationship; from  $\pm 0.7$  to  $\pm 0.9$  a high relationship, and correlations in the range from  $\pm 0.9$  to  $\pm 1.0$  a very high relationship between the characteristics. If a high or very high correlation exists between two characteristics, they are probably measuring different aspects of the same thing, and could perhaps be combined in some way to obtain a single ratio.

It will be seen from Tables 69 and 70 that the expert has chosen the characteristics well. In the case of both the *Acaena*

---

<sup>1</sup>A.L. Kidd, 'Human factors in expert systems', in Coombes, K., (Ed.), *Proceedings of the Ergonomic Society Conference 1983*, Taylor and Francis, London, 1983, (not seen), quoted by J.G. Gammack, 'Modelling expert knowledge using cognitively compatible structures', in *Third International Expert Systems Conference*, Learned Information (Europe) Ltd, London, 1987, p. 192.

<sup>2</sup>This is rarely achieved in the case of botanical specimens.

and *Danthonia* data there are no very high correlations. The *Acaena* data has the low result of 1.5% high correlations, and the *Danthonia* data only 1% high correlations.

Where there are high correlations between characteristics, there is potential to experiment by using ratios of them to obtain dimensionless measures.<sup>1</sup> Dimensionless measures intrinsic to each taxa, but varying between the taxa, would seem to have the potential to greatly assist the identification of taxa regardless of the different stages of plant growth. If characteristics exhibiting high correlations over all the taxa are used, the chance of misclassification would be lessened, (compared with a dimensionless measure obtained from characteristics with lower or null correlations).

Some ratios had already been included in the *Danthonia* data, (although they had been chosen "by eye", without the benefit of a correlation analysis). It is interesting to notice that there is a high correlation (0.75) between the ratios "Ratio of 'length of awn' to 'length of body of lemma'" and "Ratio of 'length of lateral lobe' to 'length of body of lemma'", suggesting they may both estimate some sort of dimensionless constant intrinsic to the species. However the high correlation is due to the highest correlation found in either data, (0.897) observed between the characteristics "Lemma, length of awn" and "Lemma, length of lateral lobes" existing across all species of the *Danthonia* specimens. Both of these characteristics are also correlated highly (0.86 and 0.79 respectively) with the characteristic "Glumes, length".

It will be noted that the most highly correlated *Acaena* characteristics (0.86) are "Width of Leaflets" and "Length of Leaflets"; perhaps a combination worth trying would be a length/breadth ratio.

---

<sup>1</sup>Matheus comments that one of the purposes of the constructive induction promoted by himself and Rendell is to be able to meet the need that 'a general learning system needs to be capable of generating appropriate new features as required'. If features such as the ratios suggested above could be generated, constructive induction could prove useful in the construction of keys for the identification of botanical specimens. This is an avenue for future investigation. See Matheus, Christopher, 'A Constructive Induction Framework', in Segre, Alberto Maria (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishing Inc., San Mateo, U.S.A., 1989, p. 474.

This is an area that would seem to merit further investigation<sup>1</sup>.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
Length of short spines	V1	1													
Number of short spines	V2	0.32	1												
Length of long spines	V3	0.38	-0.3	1											
Number of long spines	V4	-0.1	-0.5	0.63	1										
Amount of hairs on spine	V5	0.1	-0.3	0.38	0.58	1									
Degree to which spines are unequal	V6	-0.2	-0.6	0.5	0.83	0.51	1								
Degree of hairiness of fruit	V7	0.01	-0.2	0.31	0.42	0.58	0.36	1							
Roundness/angularity of fruit	V8	-0.2	0.42	-0.4	-0.6	-0.6	-0.3	1							
Length of style	V9	0.02	-0.3	0.22	0.27	0.28	0.34	0.34	-0.3	1					
Number of styles	V10	0.08	0.07	0.11	0	0.07	0	0.09	0.03	0.05	1				
Length of stamen	V11	-0.2	-0.1	-0.1	-0.1	-0.1	0.01	0.05	0.03	0.33	0.08	1			
Number of stamens	V12	-0.1	0.05	0.04	0	0.09	0.01	0.1	0	0.26	0.06	0.46	1		
Width of sepal	V13	0.1	-0.1	0.06	0	0.05	0.04	0.15	0	0.34	0	0.44	0.33	1	
Length of sepal	V14	0.03	-0.2	0.12	0.1	0.12	0.14	0.16	-0.2	0.42	0.05	0.62	0.38	0.71	1
Degree of hairiness of sepals	V15	-0.1	0.29	-0.2	-0.4	-0.2	-0.4	0.01	0.35	-0.1	0.03	0.03	0.05	0	-0.1
branched/unbranched inflorescence	V16	0.09	-0.1	0.1	0.13	0.02	0.16	-0.1	-0.1	-0.1	-0.1	-0.4	-0.3	-0.2	-0.3
Width of leaflets	V17	0.1	0.13	-0.1	-0.1	0.06	-0.1	0.13	0	0.21	0.04	0.17	0.09	0.19	0.23
Length of leaflets	V18	0.14	0.22	-0.1	-0.1	0	-0.2	0.05	0.03	0.2	0	0.07	0.03	0.18	0.16
Leaflet length/serration depth ratio	V19	0.04	-0.1	0.21	0.09	0.04	0.15	0.05	-0.1	0	0.04	0.02	0.13	0.17	0.19
Number of serrations on each leaflet	V20	0.2	0.39	-0.3	-0.3	-0.1	-0.4	0	0.16	0.05	-0.1	0.04	0.03	0.11	0.14
Most hairs on vein/on bottom of leaf	V21	0.09	0.52	-0.3	-0.7	-0.5	-0.7	-0.3	0.57	-0.2	0	0.07	0	0.02	-0.1
Hairiness of top of leaf	V22	0.05	0.44	-0.3	-0.5	-0.4	-0.5	-0.2	0.49	-0.2	-0.1	0.06	0	0.05	-0.1
Number leaflets	V23	0.03	0.37	-0.3	-0.5	-0.4	-0.5	-0.2	0.48	-0.2	-0.1	0.3	0.14	0.26	0.15
Stipule width	V24	0	0.21	-0.2	-0.2	0	-0.2	0.05	0.14	0.04	-0.1	0.34	0.27	0.23	0.25
Stipule length	V25	0.06	0.11	-0.1	-0.2	-0.1	-0.2	0.02	0.11	0.12	-0.1	0.35	0.16	0.28	0.34
Orientation of hairs on petiole	V26	0	-0.4	0.34	0.52	0.47	0.55	0.26	-0.5	0.31	0.05	0	0.05	0.07	0.19
Density of hairs on petiole	V27	-0.2	0.33	-0.3	-0.6	-0.5	-0.5	-0.3	0.61	-0.3	0	0.05	0	-0.1	-0.1
Leaf length	V28	0.15	0.08	0.02	0.1	0.13	0.03	0.12	-0.1	0.07	0	0	0.02	0.1	0.04
Orientation of hairs on scape	V29	-0.3	-0.1	-0.2	-0.2	-0.2	-0.1	-0.2	0.31	-0.2	0	0	0	-0.1	-0.1
Density of hairs on scape	V30	-0.1	0.3	-0.4	-0.6	-0.5	-0.6	-0.3	0.54	-0.3	0.01	0.02	0	0.07	-0.1
Length Scape	V31	0.4	0.22	0.07	0.11	0.18	0	0.14	-0.1	0.2	0	0.03	0.11	0.14	0.17

Table 69 - Correlations between *Acæna* characteristics - Part 1

<sup>1</sup>Some work had been done in this area, but will not be referred to here, as it is outside the scope of this thesis.



	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	V31	
branched/unbranched inflorescence	V16	1															
Width of leaflets	V17	-0.2	1														
Length of leaflets	V18	-0.2	0.86	1													
Leaflet length/serration depth ratio	V19	-0.1	-0.1	0	1												
Number of serrations on each leaflet	V20	-0.1	0.4	0.49	-0.2	1											
Most hairs on vein/on bottom of leaf	V21	-0.1	0.09	0.22	-0.1	0.42	1										
Hairiness of top of leaf	V22	-0.2	0	0.04	0.09	0.3	0.66	1									
Number leaflets	V23	-0.2	0.14	0.21	-0.1	0.4	0.63	0.48	1								
Stipule width	V24	-0.4	0.25	0.24	0	0.26	0.2	0.13	0.32	1							
Stipule length	V25	-0.3	0.39	0.39	0	0.3	0.17	0.06	0.36	0.64	1						
Orientation of hairs on petiole	V26	0.1	-0.1	-0.2	0.13	-0.3	-0.6	-0.5	-0.5	-0.1	-0.2	1					
Density of hairs on petiole	V27	-0.1	-0.2	-0.1	0.01	0.21	0.62	0.64	0.44	0.12	0.08	-0.6	1				
Leaf length	V28	0	0.38	0.38	0.05	0.2	0	-0.1	0.08	0.08	0.1	0.04	-0.2	1			
Orientation of hairs on scape	V29	0.02	-0.1	-0.2	0	-0.1	0.15	0.2	0.13	0	0	-0.2	0.27	-0.2	1		
Density of hairs on scape	V30	-0.2	0	0.08	0	0.24	0.63	0.59	0.52	0.11	0.15	-0.7	0.77	-0.1	0.32	1	
Length Scape	V31	-0.1	0.58	0.57	0.02	0.37	0.03	-0.1	0.09	0.25	0.26	0	-0.2	0.35	-0.2	-0.2	1

Table 69 - Correlations between *Acæna* characteristics - Part 2

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Height of Culm	V1	1								
Leaf sheaf, glabrous/pilose	V2	0.18	1							
Leaf sheaf, shining/dull	V3	0.08	0.08	1						
Leaf sheaf, prominently ribbed/smooth	V4	-0.2	-0.1	0.05	1					
Ligule, length of cilia	V5	0.14	0.08	-0.1	0	1				
Ligule, marginal tuft, length hairs	V6	0.12	0.32	-0.1	-0.1	0.29	1			
Ligule, marginal tuft, number of hairs	V7	0.15	0.18	0	-0.2	0.02	0.33	1		
Blade flat/inrolled/infolded/rolled	V8	0	-0.1	0.03	0.18	0.12	0	-0.1	1	
Blade pilose/glabrous	V9	-0.2	-0.4	0.12	0.13	-0.1	-0.3	-0.3	0.07	1
Culm, number of nodes	V10	0.46	0.17	0.05	-0.1	0.14	0.27	0.04	-0.1	-0.3
Culm, glabrous/scabrous/pilose below the panicle	V11	0.04	0.06	0.07	-0.1	-0.1	-0.1	0.14	-0.1	-0.1
Panicle, length	V12	0.72	0.19	0.04	-0.2	0.23	0.18	0.13	-0.1	-0.1
Panicle, ratio 'base to broadest width' to 'total length'	V13	0	-0.1	0	-0.1	-0.3	-0.3	0.16	-0.1	0.07
Panicle, approx. number spikelets	V14	0.45	0.09	0.07	-0.1	0.19	0.22	0.07	0	0.06
Panicle, glabrous/pilose branches	V15	-0.1	0.03	0	0.01	0.07	0.02	0.05	0.01	0
Spikelet, number of florets	V16	0.32	0.03	0.06	-0.1	0.04	-0.1	0.07	-0.1	-0.2
Spikelet, ratio floret to glume size	V17	0.09	-0.1	0.15	-0.1	-0.4	-0.2	0.41	-0.2	0
Spikelet, ratio of length of awn exerted to total awn length	V18	0.19	0.13	0	-0.2	0.04	0.14	0.23	-0.2	-0.2
Glumes, length	V19	0.52	0.13	-0.1	-0.2	0.3	0.21	0.14	-0.1	-0.2
Glumes, breadth	V20	0.16	0.06	-0.2	-0.1	0.39	0.25	-0.1	0.1	-0.2
Glumes, number of nerves extending more than half way to apex	V21	0.04	0.13	-0.1	0.04	0.23	0.2	-0.1	0.06	-0.2
Glumes, acuminate/truncate	V22	0	0	0.03	0	-0.1	-0.1	0	0	-0.1
Lemma, length of body	V23	0.37	0.23	-0.2	-0.2	0.4	0.36	0.06	-0.1	-0.3
Lemma, length of callus hairs	V24	0.33	0.16	-0.2	-0.1	0.33	0.27	0.08	0	-0.3
Lemma, length of the lower row of hairs	V25	0.22	0.12	-0.2	-0.1	0.41	0.31	0.11	0.04	-0.4
Lemma, number of tufts of hairs in lower row	V26	0.06	0.11	0.02	-0.1	0.25	0.18	0.07	0.05	-0.1
Lemma, length of the upper row of hairs	V27	0.23	-0.1	-0.1	0	0.34	0.3	0	0.09	-0.1
Lemma, number of tufts of hairs in upper row	V28	0	0.15	0.08	0	0.13	0.29	0.11	0	-0.1
Lemma, ratio 'distance from base to upper row hairs' to 'lemma body length'	V29	0.16	0.05	0.12	0.01	0.05	-0.1	-0.2	0.06	0.1
Lemma, number of nerves extending more than half way to apex.	V30	0.28	0.1	0	-0.3	0.05	0.05	0.18	-0.1	-0.2
Lemma, length of lateral lobes	V31	0.43	0.24	-0.2	-0.2	0.34	0.33	0.08	-0.1	-0.3
Lemma, ratio of 'flat length' to 'total length' of lateral lobe.	V32	-0.3	-0.1	0.06	0.11	-0.1	0	-0.1	0.12	0.07
Lemma, length of awn	V33	0.51	0.22	-0.2	-0.2	0.4	0.33	0.08	-0.1	-0.2
Lemma, ratio 'length of column' to 'length of bristle' of awn	V34	-0.2	-0.2	0.07	-0.1	-0.3	-0.3	0.15	-0.1	0.11
Lemma, length of callus	V35	0.34	0.16	-0.1	-0.1	0.29	0.21	0.07	0	-0.3
Ratio of 'distance of base to widest point' to 'total length' of palea.	V36	-0.1	-0.1	0.04	0.1	0.03	0	0.21	0.01	-0.1
Palea, rounded/bifid/acuminate apex	V37	-0.2	0	-0.1	0.05	0.01	0	0.01	0	-0.2
Palea, ratio of 'length of palea' to 'length of body of lemma'	V38	-0.1	-0.2	0.18	0.06	-0.3	-0.2	-0.1	0.06	0.4
Ratio of 'length of lateral lobe' to 'length of body' of lemma.	V39	0.28	0.15	-0.1	-0.1	0.19	0.15	0.07	0.01	-0.1
Ratio of 'length of awn' to 'length of body of lemma'	V40	0.27	0.11	0	-0.1	0.19	0.09	0.09	0.11	-0.1
Ratio of 'length of awn' to 'length of lateral lobe of lemma'	V41	0.01	0.1	0.02	0.01	0.01	0.1	0.12	0.1	0.12

	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	
Culm, glabrous/scabrous/pilose below the panicle	V11	1									
Panicle, length	V12	0.08	1								
Panicle, ratio 'base to broadest width' to 'total length'	V13	0.16	0	1							
Panicle, approx. number spikelets	V14	-0.1	0.64	-0.1	1						
Panicle, glabrous/pilose branches	V15	0.03	0	0	0	1					
Spikelet, number of florets	V16	0.12	0.3	0.09	0.07	-0.1	1				
Spikelet, ratio floret to glume size	V17	0.24	0.05	0.39	-0.1	0	0.2	1			
Spikelet, ratio of length of awn exerted to total awn length	V18	0.17	0.14	0.03	-0.1	0.06	0.35	0.24	1		
Glumes, length	V19	0.08	0.64	-0.1	0.25	-0.1	0.41	-0.1	0.2	1	
Glumes, breadth	V20	-0.1	0.19	-0.3	0.03	0	0.02	-0.5	0	0.59	1
Glumes, number of nerves extending more than half way to apex	V21	-0.1	0.01	-0.3	-0.1	0.03	0.09	-0.5	0.04	0.22	0.36
Glumes, acuminate/truncate	V22	0.04	0	0.02	-0.1	0.01	0	0.03	0.01	-0.1	-0.1
Lemma, length of body	V23	0	0.37	-0.3	0	0.01	0.24	-0.3	0.27	0.69	0.69
Lemma, length of callus hairs	V24	0	0.35	-0.2	0.01	0.03	0.13	-0.2	0.17	0.63	0.64
Lemma, length of the lower row of hairs	V25	-0.1	0.27	-0.2	0.14	0.07	0.11	-0.2	0.08	0.48	0.61
Lemma, number of tufts of hairs in lower row	V26	0.04	0.2	-0.2	0.32	0.02	0	-0.1	-0.1	0.1	0
Lemma, length of the upper row of hairs	V27	-0.2	0.38	-0.3	0.4	0	-0.1	-0.3	-0.1	0.56	0.57
Lemma, number of tufts of hairs in upper row	V28	-0.1	0.04	-0.2	0.31	0.01	-0.3	-0.2	-0.2	-0.1	0.03
Lemma, ratio 'distance from base to upper row hairs' to 'lemma body length'	V29	0.04	0.26	0.01	0.16	-0.1	0.1	0	-0.1	0.16	0
Lemma, number of nerves extending more than half way to apex.	V30	0.15	0.23	0.03	0.11	0	0.16	0.31	0.11	0.33	0.1
Lemma, length of lateral lobes	V31	0.14	0.49	-0.1	0.18	0.02	0.42	-0.2	0.32	0.79	0.5
Lemma, ratio of 'flat length' to 'total length' of lateral lobe.	V32	-0.2	-0.3	-0.1	-0.1	0.04	-0.4	-0.1	-0.3	-0.4	0
Lemma, length of awn	V33	0.07	0.61	-0.1	0.25	0	0.42	-0.2	0.33	0.86	0.54
Lemma, ratio 'length of column' to 'length of bristle' of awn	V34	0.25	-0.2	0.37	-0.2	-0.1	0.05	0.53	0	-0.1	-0.2
Lemma, length of callus	V35	0.03	0.41	-0.2	0.04	0.04	0.37	-0.2	0.26	0.59	0.43
Ratio of 'distance of base to widest point' to 'total length' of palea.	V36	0.05	0	0.22	0	0.07	0.06	0.22	0.04	0.03	-0.1
Palea, rounded/bifid/acuminate apex	V37	0.02	-0.3	0.06	-0.2	0.01	0.08	-0.1	0	-0.1	0.05
Palea, ratio of 'length of palea' to 'length of body of lemma'	V38	0.05	0.03	0.21	0.1	-0.1	-0.2	0.35	-0.2	-0.2	-0.3
Ratio of 'length of lateral lobe' to 'length of body' of lemma.	V39	0.24	0.37	0	0.25	0.01	0.35	0	0.2	0.48	0.12
Ratio of 'length of awn' to 'length of body of lemma'	V40	0.11	0.36	0	0.27	0	0.23	0	0.15	0.35	0.07
Ratio of 'length of awn' to 'length of lateral lobe of lemma'	V41	-0.1	0.01	0.1	0.04	0	-0.1	0.01	-0.1	-0.1	-0.1

Table 70 - Correlations between Danthonia characteristics - Part 2

	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	
Glumes, number of nerves extending more than half way to apex	V21	1									
Glumes, acuminate/truncate	V22	0	1								
Lemma, length of body	V23	0.47	-0.1	1							
Lemma, length of callus hairs	V24	0.39	-0.1	0.77	1						
Lemma, length of the lower row of hairs	V25	0.32	-0.1	0.6	0.66	1					
Lemma, number of tufts of hairs in lower row	V26	0.14	0	0.01	0.03	0.22	1				
Lemma, length of the upper row of hairs	V27	0.28	-0.1	0.45	0.46	0.53	0.25	1			
Lemma, number of tufts of hairs in upper row	V28	0.19	-0.1	0	0.09	0.28	0.45	0.31	1		
Lemma, ratio 'distance from base to upper row hairs' to 'lemma body length'	V29	0.01	0.01	0	0.1	0	0.18	0.11	0.1	1	
Lemma, number of nerves extending more than half way to apex.	V30	-0.1	-0.1	0.19	0.25	0.25	0.07	0.19	-0.1	0.07	1
Lemma, length of lateral lobes	V31	0.3	-0.1	0.67	0.56	0.53	0.22	0.46	0	0.15	0.29
Lemma, ratio of 'flat length' to 'total length' of lateral lobe.	V32	0.05	0.01	-0.2	-0.1	-0.1	-0.2	0.01	0.18	-0.1	-0.2
Lemma, length of awn	V33	0.26	-0.1	0.72	0.6	0.5	0.12	0.5	-0.1	0.11	0.27
Lemma, ratio 'length of column' to 'length of bristle' of awn	V34	-0.4	0	-0.3	-0.2	-0.2	-0.2	-0.2	-0.2	-0.1	0.14
Lemma, length of callus	V35	0.31	-0.1	0.69	0.64	0.52	0.14	0.31	-0.1	0.15	0.26
Ratio of 'distance of base to widest point' to 'total length' of palea.	V36	-0.2	0.02	-0.1	0	0.12	0.05	0.04	0.01	0	0.18
Palea, rounded/bifid/acuminate apex	V37	0.13	0.01	0.03	0.06	0.12	0.04	-0.1	0.07	-0.1	0
Palea, ratio of 'length of palea' to 'length of body of lemma'	V38	-0.4	0.05	-0.5	-0.4	-0.4	-0.1	-0.1	-0.2	0.13	-0.1
Ratio of 'length of lateral lobe' to 'length of body' of lemma.	V39	0	0	0.04	0.15	0.3	0.34	0.23	0	0.16	0.28
Ratio of 'length of awn' to 'length of body of lemma'	V40	-0.1	0	-0.1	0.08	0.16	0.15	0.14	-0.1	0.06	0.2
Ratio of 'length of awn' to 'length of lateral lobe of lemma'	V41	-0.1	0.03	-0.2	-0.1	-0.2	-0.2	-0.1	-0.2	-0.2	-0.1

Table 70 -Correlations between Danthonia characteristics - Part 3

	V31	V32	V33	V34	V35	V36	V37	V38	V39	V40	V41	
Lemma, length of lateral lobes	V31	1										
Lemma, ratio of 'flat length' to 'total length' of lateral lobe.	V32	-0.6	1									
Lemma, length of awn	V33	0.9	-0.5	1								
Lemma, ratio 'length of column' to 'length of bristle' of awn	V34	-0.2	0	-0.2	1							
Lemma, length of callus	V35	0.67	-0.3	0.65	-0.3	1						
Ratio of 'distance of base to widest point' to 'total length' of palea.	V36	0.02	0	0.04	0.15	0.01	1					
Palea, rounded/bifid/acuminate apex	V37	0.05	-0.1	0	0.07	0.07	0.11	1				
Palea, ratio of 'length of palea' to 'length of body of lemma'	V38	-0.4	0.21	-0.4	0.31	-0.5	0	-0.3	1			
Ratio of 'length of lateral lobe' to 'length of body' of lemma.	V39	0.71	-0.5	0.57	0	0.34	0.07	0.06	-0.1	1		
Ratio of 'length of awn' to 'length of body of lemma'	V40	0.36	-0.3	0.47	0	0.18	0.1	0	-0.1	0.75	1	
Ratio of 'length of awn' to 'length of lateral lobe of lemma'	V41	-0.4	0.26	-0.1	0.02	-0.2	0.06	-0.1	0.02	-0.4	0.2	1

Table 70 - Correlations between Danthonia characteristics - Part 4

## E.4 Consistency of Data

Since the Selecta-key method uses different methodology for distributions where the assumption that the distributions are mesokurtic can be assumed to apply, and for distributions where this assumption may be rejected, the form of the data is of interest.<sup>1</sup> The results of tests on the form of the distributions of measurements of data characteristics are discussed in section E.4.1. Both Selecta-key and some of the other methods with which it is compared are affected by the presence in the data of outliers. The results of outlier checks are discussed in section E.4.2.

### E.4.1 Form of the data

Tests were performed for each characteristic of each taxa to see if the null hypothesis 'that there is no difference between the distribution of the characteristic's statistic and a set of data of similar size drawn randomly from a normal distribution' could be rejected.<sup>2</sup>

For the *Acaena* data, the null hypothesis was rejected in 66% of the cases. Of this 66%, 9% were rejected because there was an inadequate number of observations for the test used to be considered valid, 9% because the characteristic was a discrete characteristic, and the remaining 48% were rejected because the distribution was not sufficiently close to a normal distribution.

Since it was also planned to examine the data for any natural clustering, (to see if this coincided in any way with the species distribution) the kurtosis of the distribution was also of interest.<sup>3</sup>

---

<sup>1</sup>A leptokurtic distribution is one in which the observations are clustered too tightly about the mean to be considered a normal distribution, (i.e. the distribution is too "peaky"). A platykurtic distribution is the opposite, the observations being clustered too loosely about the mean, (i.e. the distribution is too "flat"). A mesokurtic distribution is another name for a distribution which meets the statistical requirements of a normal distribution.

<sup>2</sup>The Shapiro-Wilks statistic was used for this purpose, low values of the statistic leading to rejection of the null hypothesis. If there were less than three observations, no test was used, and the null hypothesis rejected. If there were more than 6 observations, the significance level of the Shapiro-Wilks statistic was obtained by use of Royston's approximate normalising transformation.

<sup>3</sup>See SAS Institute Inc, *SAS/STAT User's Guide, Release 6.03*, SAS Institute Inc., Cary, NC, 1988, p. 80, for precautions (relating to clustering methodology) to be taken regarding the null hypothesis that the sample under consideration is a

The rejected 48% was made up of 23% leptokurtic distributions, 17% platykurtic distributions, and in the remaining 8% there were not sufficient valid observations to allow an estimate of kurtosis to be made.

In the case of the *Danthonia* data, the null hypothesis was rejected in 69% of the cases. Of this 69%, none were rejected because there was an inadequate number of observations, 24% because the characteristic was a discrete characteristic,<sup>1</sup> 20% because the characteristic was effectively discrete<sup>2</sup>, and the remaining 25% were rejected because the distribution was not sufficiently close to a normal distribution. The rejected 25% was made up of 14% leptokurtic distributions, and 11% platykurtic distributions.<sup>3</sup>

#### E.4.2 Presence of Outliers

Next the data was examined for outliers.<sup>4</sup> This was done by noting all the data in each characteristic for each taxa which fell outside 'x' sigma limits for that distribution, (assuming the distribution was mesokurtic).<sup>5</sup>

Depending on the standard applied, somewhere between about 10% and 18% of the *Acaena* data appeared to deserve a

---

random sample drawn from a mesokurtic distribution, particularly if the data is sampled from a distribution which is platykurtic.

<sup>1</sup>The methodology of applying a test of this kind to a distribution which takes discrete values may be questioned; however in some circumstances the distribution of a discrete variable may be such that it is acceptable. While this is unlikely in this circumstance, it was felt that the test should be applied on the off-chance that one of these cases did satisfy the requisite conditions.

<sup>2</sup>These included several ratios which had been calculated to only one significant figure, and the narrowness of their range made them effectively a discrete variable.

<sup>3</sup>A  $\chi^2$  test was applied to the 14% and 11%, and the null hypothesis that there was no difference between these values and an expected value of 50% of each type of distribution could not be rejected at the 0.05 level.

<sup>4</sup>A "outlier" is an observation or specimen which is significantly different from the rest of the group of observations to which it supposedly belongs.

<sup>5</sup>Where 'x' was in the range 2 to 9, plus >9. Since this was an indicative statistic, rather than an exact test, it was felt that the normal assumption would not be too limiting. Note also that, when examining the outliers, allowance has to be made that (e.g.) 5% of the data can be expected to fall outside the  $\pm 2s$  limits in the case of a normal distribution anyway, so the outlier may be just part of a normal distribution.

second look.<sup>1</sup> However this figure was markedly complicated by the number of incomplete observations.

The *Danthonia* data was more complete. Interestingly, instead of the expected 5% of characteristic observations outside the  $\pm 2\sigma$  limits, there were only 3.6%.<sup>2</sup> However many of these were clustered on possibly anomalous specimens, and thus there were specimens which could be considered outliers. The *Danthonia* outlier figures for specimens possibly deserving a second look were about half the percentages calculated for the *Acaena* data.<sup>3</sup>

An outlier could have come from two main sources; a data entry error, or some anomaly associated with the specimen being measured. Section E.4.2.1 discusses the results of examining the data looking for possible data entry errors. Section E.4.2.2 discusses the results of examining the data looking for possibly anomalous specimens.

#### *E.4.2.1 Examination for possible Data Entry errors.*

Considering the two main sources of outliers, this examination was theoretically the easiest to make; one merely had to carefully compare a print-out of the computer data with the original version of the data. However a post-hoc examination was the best that could be done, because the original data was not available.<sup>4</sup>

---

<sup>1</sup>The SAS/STAT User's Guide (SAS Institute Inc, 1988) suggests that, in practice, about 10% of the data may be outliers. The range of possible outlier percentages observed in this examination of the data are approximately allied to this figure; the exact estimate depending on the "standard" for an outlier. The upper figure of 18% would represent a fairly extreme definition of an "outlier", and the lower figure would be more in line with the SAS/STAT implied "standard".

<sup>2</sup>Strictly, 5% of the observations fall outside  $\pm 1.96$  (not  $\pm 2$ ) standard deviations. However since these investigations can be characterised as, at best, approximate, '1.96' is rounded to '2' in this discussion.

<sup>3</sup>It is recommended that this type of data examination be considered whenever new data is obtained, particularly if the tests can be carried out whilst the person who made the observations can still remember the circumstances under which the collection took place. Whilst "correct" data is not necessary if (as in this case) one's sole purpose is to conduct comparative tests of inductive learning algorithms, it would be obviously be preferable to start from good data if one wished to (e.g.) publish a definitive key for the taxa being examined.

<sup>4</sup>The situation where the original data is not available may not be as uncommon in the future as would seem probable at first thought, given that sample data sets may increasingly have been obtained over a network via ftp or a similar data transfer protocol. In cases like this a data analysis of the type carried out in this

To do this examination, tests for outliers were run, and each outlier specimen was examined in the light of this information.

Some types of data entry errors are more easily detected than others. One type of data entry error would be to tap a key too lightly, resulting in the character not being entered into the machine. A possible example of this type occurs in the "Number of leaflets" characteristic of the taxa *Acaena agnipila* var. *aequispina*, where data values which were otherwise in the range of 16 to 21, included values of 10.5 and 2.0. The latter was more than four standard deviations from the mean, and would fit the possibility of a mis-keyed 20.0, with the first zero not registering. Another possibility of the same sort occurred in the "length of leaf" characteristic of *Acaena echinata* var. *retrorsumpilosa*, where a range of observations of values 3.5 to 25 included a value of 120 (more than 6 standard deviation from the mean), perhaps a mis-keying of the decimal point of 12.0.<sup>1</sup>

A second possibility is hitting an adjoining key by mistake. The reading of 10.5 mentioned in the previous paragraph could be an example of this, (a mis-keying resulting in 10.5, not 20.5). Similarly, with every other reading being 3, a value of 4 (seven standard deviation from the mean) in an observation of a characteristic of *Acaena echinata* var. *retrorsumpilosa* is at least worth checking. Since some of this data was entered using a numeric keypad, an error caused by pressing the key above or below is also a possibility. Measurements of the "length of the Glumes" of specimens of *Danthonia Pauciflora* fall almost completely in the range of the high fives to low sixes, and it seems tempting to suspect that a data point of 9.0 (4 standard deviation from the mean) may be an example of this type of mis-keying, (9 being directly above 6 on the numeric keypad).

---

Appendix may well be a useful first step in the examination of a data set the experimenter has not personally gathered.

<sup>1</sup>In an industrial situation, applying the usual standards of quality control, an item with this degree of deviation would be rejected as unsatisfactory. However caution must be displayed, as botanic specimens tend to exhibit a far greater range of variation than generally occurs in the case of industrial or psychological measurements. P.A. Collier (private communication) instanced his observation of some *Danthonia* specimens growing in a cemetery, where a plant growing in the shade of a tombstone grew large and "leggy" as it reached for the sunlight, it being an order of magnitude bigger than other specimens growing in full sunlight nearby. *Danthonia* also responds well to appropriately-fertilised sites.



Another possibility is the keyboard operator mis-reading the measurer's handwriting. Because a carelessly written zero can resemble a 6, it is tempting to suspect a mis-read zero in the case of a 6 appearing in the "Number of Longspines" characteristic of *Acaena agnipila* var. *tenuispica*. Every other reading is zero, and the single 6 in the data is more than 4 standard deviations from the norm. Similarly a badly written 3 may be mistakenly read by the data entry operator as an 8. Again one is tempted to speculate that this may have been the case in an observation of the "length of the long spines" of a specimen of *Acaena echinata* var. *retrorsumpilosa* where the mean is 3, and the single reading of 8.0 is 5 standard deviations from the mean.

If the original data (from which the computer copy was keyed) was available, these suspicions could be rapidly resolved. In this case the original data was not available, so the uncertainties remain. The uncertainties will not effect the comparison between different key-producing algorithms being made in this thesis, but would be relevant if a definitive identification key for either species was to be produced, (this being the purpose of the key-producing algorithms).

#### *E.4.2.2 Examination for possibly anomalous specimens*

The second reason for the occurrence of outliers is the measurement of anomalous specimens. These may be genuine specimens of the species or taxa being examined which are anomalous for a variety of possible reasons. Causes of anomaly of botanic specimens are legion; possibly including the effect of ideal or marginal growing conditions (e.g. presence or absence of natural fertilisers, selective absence or presence of necessary trace elements), local shelter from or exposure to environmental stress (e.g. exposure to wind, rain, frosts, snow, animal or insect predation or infestation), and possible sub-species variation in specimens bought about by the genetic presence of rare recessive (possibly homozygous) alleles. The relative rarity of specimens of the latter type may also lead to possible misclassification, or even re-classification as a new taxa, so examination of anomalous specimens is usually a worth-while exercise.

The results of the outlier tests used in section E.4.2.1 were re-examined with anomalous specimens in mind. The occurrence of possibly anomalous specimens may be indicated by clusters of multiple outliers in the characteristic measurements of single specimens.

When examining specimens, it is not considered unusual for some to have an observation more than two standard deviations from the norm. This could be expected to occur approximately once in 20 observations in a normal distribution. However if (e.g.) three observations of this sort occurred in the same specimen; this would suggest that a second look at the specimen may be worth while, as this type of result would occur naturally only once in approximately 8000 observations.<sup>1</sup> Some specimens in the data appear more deviant than this, e.g. of the 41 characteristics measured for one specimen of *Danthonia Pauciflora*, 9 observations are "deviant", (one 4 standard deviations from the mean, two 3 standard deviations from the mean, and six 2 standard deviations from the mean). Another example is one specimen of *Acaena agnipila* var. *agnipila* which has seven "deviant" observations (including one 3 standard deviation from the norm). Again a specimen of *Acaena agnipila* var. *tenuispica* has five (including three which are 3, and one which is 4 standard deviations from the norm). Superficially, they would seem anomalous, and either the possibility of mis-measurement or mis-classification, or the possibility of sampling over a geographic range which was not truly representative of the taxa, would appear to be worth considering.

However one must be careful, as the calculation of odds depends on the distributions of the characteristics observed being independent of one another. This is not always the case. For example, one observation of *Acaena agnipila* var. *aequispina* has two "deviant" observations, one in the "Length of Short Spines" and the other in the "Length of Long Spines" characteristic. In this case the odds would seem to be approximately 1 in 400, (as each observation is reported as being more than 2 standard deviations from the mean). However an examination of the data shows that the measurer, when

---

<sup>1</sup>If the characteristics are independent.

confronted with a specimen which had no long spines, has assigned the same length to both the "Length of Short Spines" and "Length of Long Spines" characteristics.<sup>1</sup> None of the *Acaena agnipila* var. *aequispina* specimens exhibit any long spines, thus the correlation between the two characteristics is 1.0. This means the overall chance of this specimen having these measurements is not 1 in 400 but 1 in 20, and hence this specimen is probably not exceptional, but merely a "normal" chance variation.

However even allowing for this type of correlation, there were cases in which an examination of the specimen's data would strongly suggest that a check was warranted. Because of the wide variation found in botanic specimens, checking the measurements with the measured specimen would be far preferable to immediately rejecting the measurement as being non-representative of the taxa. If the specimen is correctly identified and correctly measured, it is representative, and no matter how unusual the specimen is regarded as being, and should be included in the analysis.

## E.5 Summary

After examination of the *Acaena* and *Danthonia* data, it was concluded that, in the case of both data sets:-

- a) The characteristics measured met the requirements that they employ similar knowledge structures to those employed by the experts who might use a key prepared from the data;
- b) The characteristics chosen to be measured were, in the vast majority of cases, reasonably statistically independent;
- c) The statistical form of the sets of measurements of the characteristics were not mesokurtic (normal or Gaussian) in about  $\frac{2}{3}$  of the cases;
- d) A small number of outliers were present in forms which suggested the possibility of either data-entry errors or

---

<sup>1</sup>The measurer had not documented this, and it was only noticed as a result of this analysis.

either anomalous specimens or anomalous classification of specimens;

- e) that, considering all the above, the data sets were probably reasonably representative of the type of problems inherent in many sets of botanic data intended for classificatory purposes, and thus should prove useful data sets for use in testing classificatory methodology of systems intended for use in developing keys for use with botanic data.